

Statistical Analysis for Business Decisions

STATISTICAL ANALYSIS for Business Decisions

By

WILLIAM A. SPURR, Ph.D.

PROFESSOR OF BUSINESS STATISTICS

and

CHARLES P. BONINI, Ph.D.

ASSOCIATE PROFESSOR OF STATISTICS

*GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY*



1967

RICHARD D. IRWIN, INC.

Homewood, Illinois

311.2
S772 ✓
cap. 6

© RICHARD D. IRWIN, INC., 1954, 1961, and 1967

All rights reserved.

This book or any part thereof may not be reproduced
without the written permission of the publisher.

The previous edition of this book was published under
the title *Business and Economic Statistics*.

First Printing, June, 1967

Library of Congress Catalog Card No. 67-15848

Printed in the United States of America

PREFACE

THE ROLE of quantitative analysis in business and economics has expanded tremendously in recent years with advances in statistical theory, electronic computers, and the growing appreciation of the scientific method in general, as opposed to intuitive methods of reasoning. New analytic techniques have sprung from probability theory, operations research, and decision theory, while computers have provided an effective catalyst to their widespread adoption. The basic university courses in statistics reflect this wide diversity in subject matter, as well as the varying goals of different schools and differing levels of students.¹

It is with this great diversity in mind that we have planned this text. A broad range of topics is included, from the traditional tools of analysis to the modern concepts of simulation and Bayesian decision theory; from simple graphic techniques to sophisticated topics such as survey sampling and probability models. The instructor can structure his course by selecting subjects appropriate to the background and abilities of his students.

Since the book is planned for the general student who needs to use statistics in his chosen field of work, the principal emphasis is placed on the use of statistical methods as scientific tools in the analysis of practical business and economic problems, rather than on theory or mathematical derivations. The material has been presented as simply as possible, with a minimum of statistical jargon.

The main text requires no knowledge of mathematics beyond elementary algebra. The more advanced topics are marked by asterisks in the Table of Contents, so that the instructor in the elementary course can easily omit them if desired. Optional material—some of it involving calculus or matrix algebra—appears in the appendixes of several chapters. Some 400 problems have been included to allow flexibility in assignments and a broad range of practical applications for class discussion, home study, or laboratory work. Almost all of the text and problems have been tested in the basic statistics courses at the Stanford Graduate School of Business, and revised on the basis of student evaluation.

¹ See *The Teaching of Statistics in Business Schools*, by E. Cox, W. Spurr, and W. Peters (Washington, D.C.: American Statistical Association, 1964).

The book is divided into six parts:

1. An introduction to the basic tools of analysis, in Chapters 1–6. We believe it desirable to discuss how to find the facts, and how to present the results in tables and charts; but the instructor who wishes to move directly to the analysis of data can skip Chapters 2 and 3 (except ratio charts) without loss of continuity.

2. The elements of probability theory and the principal probability distributions are described and applied to decision-making in Chapters 7–10. Probabilities of events, payoff tables, expected values, the value of information, and decision trees are all elements of a rational procedure for making decisions under uncertainty.

3. In order to draw inferences about sample information, it is desirable to set confidence limits or test hypotheses, as described in Chapters 11–13. In practical surveys, however, simple random sampling will not usually suffice, so Chapter 14 explores a variety of other sample designs that are more efficient or practicable. This topic is too often ignored in elementary books.

4. Probabilities and sample evidence are combined through Bayes' Theorem in Chapters 15 and 16 to improve the decision-making process. Here, as in Chapters 9 and 10, economic costs and profits are explicitly included in the analysis. This topic represents an important extension of the traditional interpretation of sample information. Simulation and other recently-developed probability models are applied to business problems in Chapter 17.

5. Statistical analysis in business and economics requires considerable emphasis on time series, since the economist is vitally concerned with measuring and projecting economic growth, seasonal movements or business cycles. We therefore survey index numbers and time series analysis and forecasting, together with computer applications, in Chapters 18–21.

6. Correlation and regression techniques are widely used and misused. The reader may well wish to be content with simple regression, but multiple regression is a more powerful tool, and is easily manageable in the new computer programs, so the entire treatment in Chapters 22–24 is recommended, if time permits. Finally, we present quality control in Chapter 25 as a practical application of the theory of testing hypotheses.

The book contains enough material for a two-semester course in statistics—say Chapters 1–14 for the first term and Chapters 15–25 for the second term. It may also be used for either a one-semester course or a more advanced course, by appropriate selection of topics. For example,

a course along classical lines might be fashioned from Chapters 1–8, 11–13, and 18–22. In addition, Chapter 9, and the first parts of Chapters 10 and 15 might be included (or substituted for other chapters) if an introduction to Bayesian decision theory is desired.

An advanced course might include Chapters 7–10, 14–17, and 23–25. Other combinations of chapters may be selected to meet the requirements of specific schools and groups of students.

The authors are much indebted to Lester S. Kellogg and John H. Smith, whose major contributions to Spurr, Kellogg, and Smith, *Business and Economic Statistics* (1st ed. 1954, rev. ed. 1961; Homewood, Ill.: Richard D. Irwin, Inc.), provided the basis from which the present Chapters 1–6 and 18–19 have evolved. The general treatment of decision theory given in Chapters 9–10 and 15–16 follows in the tradition of the excellent pioneering work of Robert Schlaifer, *Probability and Statistics for Business Decisions* (New York: McGraw-Hill Book Co., Inc., 1959). The authors are also indebted to the following professors who contributed important sections to the more advanced chapters: Roy W. Jastram on statistical inference, Karl A. Fox and Oscar N. Serbein on correlation and regression, and Frank J. Williams and David S. Chambers on quality control. Professor Howard Raiffa provided valuable ideas in his seminar on decision theory at Stanford in 1966. Finally, we wish to acknowledge the generous support of the Stanford Graduate School of Business in providing both time and facilities for us to complete this task.

April, 1967

WILLIAM A. SPURR
CHARLES P. BONINI

TABLE OF CONTENTS

PART I. ANALYSIS OF BASIC DATA

CHAPTER	PAGE
1. STATISTICS IN BUSINESS AND ECONOMICS	1
Statistical Analysis as a Scientific Method, 1. The Role of Statistics in Decision-Making, 3. Statistics in Business, 4. Statistics in Economics, 6. Cautions in the Use of Statistical Data, 8. Selected Readings, 16.	
2. COLLECTION OF DATA	18
Use of Research Sources, 18. Collection of Original Data, 26. Selected Reading, 41.	
3. EFFECTIVE USE OF TABLES AND CHARTS	42
Statistical Tables, 42. Charts, 46. Common Types of Charts, 49. Arithmetic Line Charts, 49. Ratio Charts, 51. Bar Charts, 57. Selected Readings, 65.	
4. ANALYSIS OF DATA: RATIOS AND FREQUENCY DISTRIBUTIONS . . .	67
Ratios, 68. Frequency Distributions, 71. Charts of Frequency Distributions, 78. Cumulative Frequency Distributions, 82. Frequency Curves, 84.	
5. AVERAGES	94
The Arithmetic Mean, 94. The Median, 99. The Modified Mean, 102. The Mode, 102. The Geometric Mean, 104. Which Average to Use? 106. Characteristics of Averages, 107. Summary of Formulas, 109.	
6. DISPERSION	114
Purposes of Measuring Dispersion, 116. The Range, 117. The Quartile Deviation, 118. The Mean Deviation, 120. The Standard Deviation, 122. Relation between Measures of Dispersion, 129. Measures of Relative Dispersion, 132. Skewness, 133. Uses of Measures of Dispersion, 134. Summary of Formulas, 135. Selected Readings, 139.	

PART II. PROBABILITIES AND DECISION-MAKING

7. AN INTRODUCTION TO PROBABILITY THEORY	140
Basic Concepts, 140. Rules for Dealing with Probabilities, 146. Examples in the Use of Probabilities, 148. Probability Distributions, 153. Expected Value and Variance of Probability Distributions, 156. Appendix: Expected Value and Variance of Continuous Distributions, 158.	
8. THE BINOMIAL, POISSON, AND NORMAL DISTRIBUTIONS	166
The Binomial Distribution, 166. The Poisson Distribution, 173. The Normal Distribution, 176. Selected Readings, 188.	
9. PROBABILITIES AND DECISION-MAKING	189
Certainty versus Uncertainty, 189. Probability Models, 190. How to Make a Decision, 191. Subjective Probabilities and Decision-Making, 199. Decision Trees, 202. Risk in Decision-Making: The Utility of Money, 211. Appendix: Derivation of Utility Curves for Decision-Making under Uncertainty, 214.	

CHAPTER	PAGE
10. DECISION-MAKING UNDER UNCERTAINTY: THE VALUE OF ADDITIONAL INFORMATION	224
Opportunity Loss, 224. Expected Value of Perfect Information, 226. Linear Profit Functions,* 230. The Normal Distribution in Decision Making,* 234. Formulas, 242. Selected Readings, 247.	
PART III. DRAWING INFERENCES FROM SAMPLE	
11. INTRODUCTION TO STATISTICAL INFERENCE	249
Sampling Error and Bias, 250. Simple Random Sampling, 252. How Sample Means Are Distributed, 254. The Standard Error of the Mean, 260. Confidence Intervals, 265. How Big Should a Sample Be? 268.	
12. TESTS OF HYPOTHESES	276
Type I and Type II Errors, 281. Two-tailed Tests Versus One-tailed Tests, 287. Tests of Differences between Arithmetic Means, 289.	
13. INFERENCES INVOLVING SMALL SAMPLES AND PROPORTIONS	300
Small Samples, 300. Proportions, 303. Selected Readings, 314.	
14. SAMPLE SURVEY METHODS	316
Probability Sampling, 316. Simple Random Sampling, 317. Systematic Selection, 318. Stratified Sampling,* 319. Ratio Estimation,* 327. Cluster Sampling,* 334. Replicated Sampling, 339. Nonprobability Sampling, 340. Summary of Formulas, 344. Selected Readings, 355.	
PART IV. REVISING DECISIONS FROM SAMPLE EVIDENCE	
15. BAYES' THEOREM: REVISING PROBABILITIES	356
Prior and Posterior Probability Distributions, 357. Posterior Probabilities and Decision-Making, 363. Expected Value of Sample Information,* 367. Bayesian versus Classical Approach, 370.	
16. BAYES' THEOREM FOR NORMAL DISTRIBUTIONS*	377
Determining the Posterior Distribution, 377. Evaluation of Sampling Information, 384. Optimal Sample Size, 387. Formulas, 391. Selected Readings, 396.	
17. PROBABILITY MODELS AND DECISION-MAKING	397
A Bidding Model, 397. An Inventory Model,* 399. A Queuing Model,* 403. Simulation, 407. Appendix: The Monte Carlo Method for Continuous Distributions, 415. Selected Readings, 425.	
PART V. MEASURING AND FORECASTING BUSINESS CHANGES	
18. INDEX NUMBERS	427
Advantages of Index Numbers, 428. Kinds of Index Numbers, 428. Basic Methods of Constructing Index Numbers, 431. Tests of a Good Index Number, 438. Revisions of Index Numbers, 444. Some Important Indexes, 447. Selected Reading, 462.	

* Indicates sections that contain more advanced material.

CHAPTER	PAGE
19. TIME SERIES ANALYSIS: SECULAR TREND	463
Types of Business Fluctuations, 463. Secular Trend, 465. Methods of Measuring Trend, 473.	
20. SEASONAL VARIATION	502
Nature of Seasonality, 502. Purposes of Measuring Seasonality, 506. Methods of Measuring Seasonal Variation, 507. Use of Seasonal Indexes in Short-Term Forecasting, 523.	
21. CYCLICAL AND IRREGULAR FLUCTUATIONS	532
Reasons for Measuring Cycles, 534. How to Measure Cycles, 535. Cyclical Forecasting, 540. A Résumé of Statistical Methods in Forecasting, 545. Selected Readings, 549.	

PART VI. ANALYZING RELATIONSHIP FOR PREDICTION AND CONTROL

22. SIMPLE CORRELATION AND REGRESSION	551
Scatter Diagrams, 552. Regression Analysis, 554. Sampling and Regression Analysis, 563. Coefficient of Correlation, 571. Examples of Regression Analysis, 577. Cautions in the Use of Correlation and Regression Analysis, 579.	
23. MULTIPLE CORRELATION AND REGRESSION	589
Multiple Regression, 590. Estimation of Multiple Regression Coefficients, 593. Coefficient of Multiple Determination, 600. Statistical Inference in Multiple Regression,* 601. Interpretation of Multiple Regression Results,* 602. Cautions in the Use of Multiple Regression, 610. Appendix A: Introduction to Matrix Operations,* 620. Appendix B: Matrix Solution to Multiple Regression Analysis,* 627.	
24. CURVILINEAR AND TIME SERIES REGRESSION	631
Curvilinear Regression, 631. Correlation of Time Series, 640. Selected Readings, 657.	
25. STATISTICAL QUALITY CONTROL	659
Types of Variation in Quality, 659. Control Charts for Variables, 660. Control of Attributes, 672. Acceptance Sampling, 675. Selected Readings, 681.	

APPENDIXES

APPENDIX	PAGE
A. GLOSSARY OF SYMBOLS	685
B. LOGARITHMS	689
C. SQUARES, SQUARE ROOTS, AND RECIPROCALs 1-1,000	694
D. AREAS UNDER THE NORMAL CURVE	706
E. UNIT NORMAL LOSS FUNCTION	707

* Indicates sections that contain more advanced material.

xii TABLE OF CONTENTS

F. BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS	708
G. BINOMIAL DISTRIBUTION—CUMULATIVE TERMS	715
H. POISSON DISTRIBUTION—INDIVIDUAL TERMS	722
I. POISSON DISTRIBUTION—CUMULATIVE TERMS	724
J. VALUES OF t	726
K. SUMS OF SQUARES AND FOURTH POWERS USED IN TREND FITTING .	727
L. RANDOM NUMBERS	728

INDEX

INDEX	733
-----------------	-----



I. STATISTICS IN BUSINESS AND ECONOMICS

STATISTICS in today's business and economics includes: (1) statistical data, (2) statistical analysis, and (3) decision-making. One is valueless without the others. Numerical data and methods of analysis and decision-making are becoming increasingly important in business management and in every field of economics.

But what are statistical data? Not all numbers are statistical; logarithms, for instance, are merely abstract numbers. Statistical data are concrete numbers which represent objects—their counts or measurement. Statistics deals with numbers not merely as such but as expressions of significant relationships. It is not enough to collect and present the data, therefore; they must be carefully analyzed and interpreted as well, in order to make the best possible decisions based on the data. As Lord Kelvin put it:

When you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be.

STATISTICAL ANALYSIS AS A SCIENTIFIC METHOD

When masses of numerical information are to be analyzed, some means of summarization must be found which will reveal their major characteristics. Statistical analysis meets this need. Hence, in a broad sense, statistical analysis is a scientific method of studying quantitative data. It is a means of summarizing the essential features and relationships of the data and then generalizing from these observations to

determine broad patterns of behavior or future tendencies. Statistical analysis therefore is useful in any field of knowledge in which extensive numerical information is needed.

The social and biological sciences, in particular, require masses of facts in order to determine general behavior, because of the wide variation in individuals. In the physical sciences, on the other hand, precisely controlled laboratory experiments can be used instead, to a large extent. The physicist can estimate the speed of light by repeated trials, with a small error of measurement, whereas the market analyst who wishes to determine consumer preferences toward compact cars must deal with a sample of consumers who vary widely in their preferences. He must design a questionnaire, select an unbiased sample, and estimate the sampling error. Human and biological groups are more variable in behavior than are most physical phenomena, so their study requires a statistical approach even more than in the physical sciences. Statistical analysis is therefore the fundamental method of quantitative reasoning not only in business and economics but also in sociology, anthropology, psychology, education, medicine, public health, and biology.

Statistical theory is founded on the mathematics of probability, which provides the basis for determining not only general tendencies but also the reliability of each generalization. The whole process of reasoning from the specific to the general may be called *statistical inference*, as well as *generalization* or *induction*. The field of statistical analysis itself is also called *statistical methods* or merely *statistics*. The latter term is used here in the singular sense, as opposed to "statistics" in the plural sense, which refers only to the observed data themselves.¹ Applications of statistical analysis in a particular field may be known under other names connoting the idea of *measurement* or *research*, such as econometrics, biometrics, psychometric methods, or forest mensuration—also business research, economic research, or marketing research methods. Finally, statistics plays an important part in the newer fields of operations research, management science, and systems analysis.

The importance of the statistical approach to the solution of practical problems has gradually come to be realized during recent times. The progress in this direction is explained by several developments. Fundamentally, the tremendous growth of population, large-scale production, and trade that followed the Industrial Revolution has required the production and use of a vast volume of statistics in every sphere of social activity. Statistical knowledge has increased in quantity, quality, and

¹ Note that the word "data" is plural; the singular is "datum."

frequency. The expanding needs of government have accelerated this growth. As a result, fact-finding has become an integral part of economic progress.

Increasing public interest in and demand for social statistics rests, then, on the basic premise that the problems of society, as well as of natural science and technology, can be solved by the increase and diffusion of this especially matter-of-fact type of matter-of-fact knowledge. The whole world now seems to hold that statistics can be useful in understanding, assessing, and controlling the operations of society.²

Statisticians, too, have discovered new analytical techniques which have increased the value of statistical methods of planning and control. In particular, with the advent of the electronic computer in the last decade or so, the statistician has acquired a means of dealing quickly with vast quantities of data. The use of the computer has made statistical methods inexpensive and powerful tools for analysis.

The applied statisticians have also helped to dispel the aura of mystery which formerly surrounded the subject. This has been accomplished through a shift in teaching emphasis toward the applied side and through the publishing of textbooks and reference books which stress the simplicity of statistical application and avoid perpetuating the impression that one must be master of advanced mathematics in order to do statistical work.

THE ROLE OF STATISTICS IN DECISION-MAKING

Statistical data are collected and analyzed not only for the purpose of adding to scientific knowledge in general but also for the purpose of helping the rational man to make decisions. One of the most important functions of the business executive, the government official, or the administrator in any field is to make decisions. The function of statistics is to help decide what data are needed and how the data shall be collected, tabulated, analyzed, and interpreted in such a way as to lead to the best possible decision. Unfortunately, the complete facts are not usually available, so incomplete data, or samples, must be used. Statistics then provides methods that help the executive make the best decision on the basis of these incomplete facts. Hence, statistics has come to be defined as a group of methods for making wise decisions in the face of uncertainty.

Of course, statistical methods do not provide the only basis for decision-making. There are many intangible factors—the business “cli-

² Solomon Fabricant, “Factors in the Accumulation of Social Statistics,” *Journal of the American Statistical Association*, June 1952, p. 259.

mate," prospective government action, technological developments, or personnel relationships, for example—which make management an intuitive art rather than a science. Nevertheless, statistics provides the primary factual basis for reaching good decisions. The executive who masters the statistical approach to decision-making will narrow the range of uncertainty and increase his probability of making a correct decision.

As M. A. Girshick has said:

All branches of statistics . . . deal with the same basic problem, namely, the problem of decision making in the face of uncertainty. All decisions rules . . . must be evaluated by their consequences. These consequences are expressible in terms of risks, or more intrinsically, in terms of the probabilities of taking the various permissible actions which are induced by the experiment, decision rule, and the possible states of the system. In brief . . . not facts from figures but rather decisions from observations should become the main emphasis in elementary statistical observations.³

Faced with a business problem involving uncertainty, we can list the future events that may occur and the probability that each will happen together with the various acts or decisions that may be taken, and the consequence (e.g., cost) of each combination of a given act and a resulting event. The best decision rule is then the one that minimizes the expected total cost, allowing for the probabilities involved. We can also determine whether it is preferable to delay a decision and to obtain additional information before acting. This procedure provides the executive with a better basis for decision-making than he could have obtained from his unaided intuition.

The role of the electronic computer is becoming increasingly important in the decision-making process. The computer can make a simple decision itself (as in inventory control) or else perform extensive analyses to aid the executive in making a more complex decision. Statistical methods provide not only the data but also the techniques used by the computer in decision-making.

STATISTICS IN BUSINESS

The employment of statistical methods in the solution of business problems belongs almost exclusively to the twentieth century. At an earlier date, when practically all business enterprises were small, management was able to comprehend its problems in detail by personal contact. The increased size of concerns in the present period has required more planning and greater regimentation of operations. At the

³ *Journal of the American Statistical Association*, September 1953, p. 646.

same time, management has found it impossible to maintain personal contact with its problems. The alternative is control through the interpretation of numerical information. This chain of circumstances has led to the introduction of statistical methods of investigation as a primary aid in the performance of the function of management.

According to a study made by the Pacific Telephone and Telegraph Company:

Today, management at all levels is guided quite generally by facts obtained through analysis of records rather than upon knowledge obtained merely through personal observation and experience. . . . Through application of appropriate statistical methods, current performance may be measured, significant relationships may be studied, past experience may be analyzed and probable future trends appraised. . . .

The use of statistical methods and the performance of analytical work which is largely statistical in character—whether or not it happens to be carried on under the distinctive label of “statistics”—occupy a conspicuous place in the work of all departments of the company.⁴

Statistical analysis is thus used as a basis for the control of many operations in a company and for planning or forecasting its activities. Through the aid of statistical reports the executive can gain a summary picture of current operations which improves his factual basis for making valid decisions affecting future operations.

The principal statistical activities of a typical large and progressive firm are as follows:

1. A central economic research or statistical department operates under the guidance of an “economist” or “chief statistician.” This department analyzes general business trends and forecasts business activity, commodity prices, and other economic factors. It may coordinate the internal company statistics compiled by other departments and issue summary reports of operations to top executives. It also makes periodic comparisons of the company’s performance with that of its competitors.

2. A marketing research staff makes surveys of consumer preferences and purchasing power and forecasts probable future trends in sales. It may prepare a detailed sales budget for the coming year, broken down by individual products and by months. Finally, it has the responsibility for setting salesmen’s quotas by territories and products, based on past performance, income studies, and salesmen’s estimates.⁵

3. The production department maintains a “quality control” staff that minimizes defective output by means of statistical checks, as de-

⁴ *Statistics in the Telephone Business* (March 1, 1951).

⁵ See Frank D. Newbury, *Business Forecasting* (New York: McGraw-Hill, 1952), chaps. 1, 2, 15.

scribed in Chapter 25. It prepares forecasts of production based on sales forecasts and other criteria and checks actual production against these estimates. It also maintains an inventory control system and makes time and motion studies.

4. The controller's department combines statistical and accounting methods in making the overall budget for the coming year—including sales; material, labor, and other costs; and net profits and capital requirements. It may maintain a standard cost system for controlling costs and setting prices of products.

5. The personnel department makes statistical studies of wage rates, incentive systems, the cost of living, employment trends, labor turnover rates, accident rates, and results of employee selection procedures.

6. The investment department maintains security analysts who study individual stocks and bonds and the general outlook for the securities markets.

7. The credit department performs statistical analyses to determine how much credit to extend to each potential customer. Characteristics of those customers who have paid and those who have defaulted in the past are used for selecting future credit risks.

8. The executive department may include an "operations research" staff. This group consists of specialists, such as statisticians, mathematicians, and physicists, who apply scientific methods to the study of complex operations throughout the organization. The purpose is to provide top management with a factual basis for making policy decisions.

Some of the men and women who perform these functions are professional statisticians, but most of them have developed their knowledge of statistical analysis as an adjunct to their major specialties. In all departments of a business, personnel are concerned with the collection, classification, and presentation of statistics, even if their work requires no analysis. The general executive, too, must know some statistics as well as the basic principles of accounting, finance, business law, marketing, production management, and industrial relations in handling the various aspects of his job. He cannot depend entirely on specialists for his knowledge.

STATISTICS IN ECONOMICS

Economists and other social scientists are more concerned with conditions in the economy as a whole than with those in an individual concern, but they depend on statistics just as the business analyst does. Indeed, many of the statistical problems in economics are similar to, or identical with, those in business. Economists today are no longer content to theorize in abstract terms, citing statistics only as needed to buttress

their arguments. Instead, they utilize the excellent data now available to build a sound factual foundation for their reasoning. Some of the uses of statistics in economics are as follows:

1. Extensive statistical studies of business cycles, long-term growth, and seasonal fluctuations serve to expand our knowledge of economic instability and to modify older theories.

2. Measures of gross national product and personal income have greatly advanced overall economic analysis and opened up an entirely new field of study.

3. Statistical surveys of prices are essential in studying the theories of prices, pricing policy, and price trends, as well as their relationships to the general problem of inflation.

4. Financial statistics are basic in the fields of money and banking, short-term credit, consumer finance, and public finance.

5. Operational studies of public utilities, including the transportation and communication industries, require both statistical and legal tools of analysis. Such studies are necessary in connection with the federal and state regulation of these industries.

6. Analyses of population, land economics, and economic geography are basically statistical and geographic in their approach.

7. Studies of competition, oligopoly, and monopoly require statistical comparisons of market prices, costs, and profits of individual firms.

Statistical analysis is therefore carried on in every field of inductive economics—by individual professors, university economic research bureaus, chambers of commerce, trade associations, and such well-known research agencies as the National Bureau of Economic Research, the National Industrial Conference Board, the Twentieth Century Fund, and the Brookings Institution, to mention a few.

The most spectacular development of statistical analysis in economic research during recent years, however, has been in the federal government. As it has grown in size, the government has greatly expanded the scope of its statistical activities in every field of applied economics. Some agencies collect and publish statistics for their informational value to the public, while others compile data as a by-product of administrative or regulatory activities. Under the Full Employment Act of 1946 the President's Council of Economic Advisers and the congressional Joint Economic Committee employ many statistical indexes as guides in recommending to the President and Congress control measures designed to allay depression or inflation. Statistics has become as much a major tool of economic guidance and control by the federal government as it is an operational tool for individual concerns.

The various wars of the past half-century have tremendously stimu-

lated the government's need for statistical information in administering the great manpower and volumes of matériel involved. The pressures of war have also caused an accelerated development in statistical *methods*. Since the need for controls increases with the size of an enterprise, the federal government's wartime organization has required unprecedented numbers of statisticians in purchasing, logistics, operations research, and many other fields.

To conclude this introduction, we quote from M. J. Moroney's *Facts from Figures*:^{5a}

If you are young, then I say: Learn something about statistics as soon as you can. Don't dismiss it through ignorance or because it calls for thought. . . . If you are older and already crowned with the laurels of success, see to it that those under your wing who look to you for advice are encouraged to look into this subject. In this way you will show that your arteries are not yet hardened, and you will be able to reap the benefits without doing overmuch work yourself. Whoever you are, if your work calls for the interpretation of data, you may be able to do without statistics, but you won't do so well.

CAUTIONS IN THE USE OF STATISTICAL DATA

The beginner in statistical work is apt to have the attitude that numerical facts can be accepted without question. A few adverse experiences will usually dispel this initial trustfulness in favor of a healthy skepticism. The scientific attitude toward evidence is skeptical rather than either cynical or uncritically enthusiastic.

Many of the misuses that appear in statistical reports arise from failure of the authors to maintain a critical attitude toward their work. Even facts and statements that are true in some sense can be quoted out of context or presented in such a way that they are bound to be misinterpreted by most readers. As a result, the disillusioned have coined such slogans as: "There are three kinds of lies—lies, damn lies, and statistics," and conversely, "Figures don't lie, but liars figure." Many people use statistics as a drunkard uses a street lamp—for support rather than for illumination.

The scientific investigator must seek the *truth* above all. It is not enough to avoid outright falsehood; the investigator must be on the alert to detect possible distortion of truth. One can hardly pick up a newspaper without seeing some sensational headline based on scanty or doubtful data.

Several types of misuse are presented below. Some contain actual errors or falsification of facts, but others consist of entirely true statements taken out of context. All examples are taken from reputable publications, but many sources are omitted to avoid embarrassment.

^{5a} M. J. Moroney, *Facts from Figures* (Baltimore, Md.: Penguin Books, 1956).

Bias

Conscious or unconscious bias is very common in statistical work. It is easy to detect the conscious bias in an advertisement that quotes statistics to "prove" the superiority of a given product, while a competitor's ad quotes other statistics to "prove" the superiority of his own product. But many compilers of statistics have an ax to grind. A jewelers' association quotes figures purporting to show that the double-ring wedding has become "an accepted national custom." A labor organization claims that a consumer price index, on which wages are based, should be revised upward because it understates real costs, while an employers' association defends the index, pointing out components that overstate real costs. The source of the data must be considered, as well as the conclusions themselves.

Unconscious bias is even more insidious. Perhaps all statistical reports contain some unconscious bias, since the results of research must be interpreted by human beings, each of whom can judge only in terms of his own experience and his attitude toward the problem at hand. The investigator must disregard his preconceptions and avoid wishful thinking in order to attain an objective conclusion. If biased data must be used in the absence of better information, the nature and probable direction of the bias must be considered in interpreting the results.

Faulty Generalization

A basic error in statistical reasoning is to jump to a conclusion or generalization on the basis of too small a sample or one which is not typical of the whole population to which the conclusions are applied. This subject is of such importance that several chapters in this book are devoted to methods of selecting samples and making statistical inferences.

As an example of using too small a sample, a national magazine reported that a group of Colorado schoolteachers had been given a test in history and failed with an average grade of 67, indicating that Colorado schoolteachers generally were deficient in history. An official of the Colorado Education Association retorted that only four teachers had been given the test, of whom three made the respectable average score of 83 and the fourth only 20, bringing the average of the four down to 67.

An extreme case of using too small a sample is that of generalizing from a sample of one, or citing only a single case. Thus, a typewriter manufacturer advertises that "Tests by leading educators prove that students who use typewriters get up to 38% better grades."

Improper generalizations based on nontypical samples are more difficult to detect. Such samples may be adequate in size, but they differ from the total population in some essential characteristic; so the generalization is faulty. For example, a feature article in *Advertising Age* is entitled "Obits Show 'Average' Adman is Dead at 62," based on obituaries of 300 advertising men who died during the previous year. Perhaps the advertising game *does* kill men off young, but there may be two defects in the sample used: (1) Since many young men have entered this field in recent years, those who died during the past year were relatively young; the surviving ones who will live to a riper old age of course are not counted. (2) If advertising is a young man's game, as reputed, older men go into other fields and are counted there when they die. As an analogy, the average age at death of college students is about 20 years, but this does not indicate that college graduates die young.

Another example is a report in a business school alumni journal that the average graduate in the class of 1920 earned \$87,049 in a recent year. This figure was based on 18 returns received from a questionnaire mailed to 62 class members. Unfortunately, the average income is not typical if a larger proportion of those with higher incomes return the questionnaire than do those with lower incomes or if some respondents exaggerate their incomes, as is sometimes the case. Furthermore, if a few alumni have very high incomes, these figures greatly inflate the average.⁶

Faulty Deduction

Faulty deduction (in the sense that deductive reasoning is the opposite of inductive reasoning) occurs when a general statement is applied erroneously to a specific case.

Thus, an electric institute reported that "industry's generating capacity in December was 5.1 percent above electricity demands." This statement was doubtless true of the country as a whole, but it would be faulty deduction to apply it to a specific region which might be short in generating capacity. Regions, such as the Far West, that have grown rapidly in population were in fact short of power at this time.

Similarly, an opponent of health insurance may say that families generally spend only 5 percent of their total income on medical care—less than they spend on liquor or recreation. Nevertheless, medical care may be a crushing burden on an individual family in a particular year.

⁶ This example illustrates several misuses: (1) too small a sample, (2) nontypical sample, (3) spurious accuracy (see below), and (4) use of mean instead of median (see Chapter 5).

The common error of faulty deduction arises fundamentally from the human tendency to apply a valid general rule as if it were an invariant mechanical law. Such a proposition should be stated as a general tendency, instead, with allowance for individual differences.

Noncomparable Data

Comparisons are frequently made between two things that are not really alike. For example, certain airlines advertised that air travel was cheaper than first-class rail travel. The Southern Pacific Company in a series of advertisements entitled, "A short course in Railroading . . . for Airline executives," claimed that these figures were not comparable because (1) the one-way fares quoted did not make allowance for the greater reduction in round-trip fares on railroads, (2) the fares compared the cost of a chair on a plane with that of a bed or lower berth on a train, and (3) no allowance was made for the rails' carrying children free under five years of age and their extra baggage allowance.

A whisky manufacturer advertised that the price of his product (before taxes) had not increased appreciably during the past decade, without mentioning that the proportion of "grain neutral spirits" had been increased. Errors due to noncomparability affect price indexes generally, since the specifications of components vary from time to time.

A feature article in *Time* (October 28, 1957) praising West Germany's price stability, says: "In the U.S. the cost of living . . . has reached a record high of 121 (the 1948-49 average: 100). . . . In this global sea of inflationary troubles there is one major island where enterprise . . . has achieved a basic stability of consumers' prices. In West Germany the cost-of-living index was up a modest 16 points from 1950 levels." Here the base periods are not comparable, nor is the percentage comparison clear. Since the base period of the German index must be assumed to be 1950, its rise was 16 percent, which just about equaled that in the United States, if computed from 1950 rather than from the 1947-49 base (erroneously reported by *Time* as 1948-49). In order to make a fair comparison between two things, it is essential that they have the same pertinent characteristics.

Errors in Semantics

Slanted or colored words are sometimes used to influence the reader or listener. Witness political campaigns. One common error is the use of "leading questions" in surveys to suggest the desired answers. For example: "Why do you prefer our product?" One market analyst reports that even such a seemingly innocuous wording as, "Have you read — [the latest novel]?" brought a much larger proportion of favorable replies

than when a similar group of people was asked the question, "Do you happen to have read — [the same novel]?" Note, too, the "record high" in U.S. prices as contrasted with the "modest" rise in German prices in the *Time* article above. The impartial investigator must check his words, as well as his figures, for possible bias.

Assuming Causation from Correlation

This common fallacy of reasoning, sometimes called *non sequitur* ("it doesn't follow") by the pundits, means that because one thing precedes another in time, it is assumed to be the cause of the other. You get wet feet, then catch cold. The wet feet are then assumed to be the cause of the cold. A student writes a correspondence institute: "I am well pleased with the law course. A month after enrolling, my salary was increased in the amount of 20 percent." *Non sequitur*.

An article entitled "They Put a Parson on the Payroll" in a popular magazine states: "In just two years religion-on-the-job has accomplished several pretty wonderful things . . . labor turnover has dropped from 7.61 to 5.22% in two years, the accident rate has declined approximately 40%, and absenteeism is much lower than it used to be." The assumption that the improvements in labor conditions were due to spiritual counseling does not appear to be justified in view of the many other factors that affect labor turnover and accident rates.

Many business cycle theorists in the past have found that some particular economic factor has correlated with general business activity, and hence they have assumed that this factor is "the cause" of business cycles. Unfortunately, economic and business affairs represent a complex of interacting forces. The search for simple cause-and-effect relationships is naïve and unrealistic.

Similarly, large-scale studies have established a correlation between smoking and lung cancer. However, it is a matter of bitter dispute whether heavy smoking *causes* lung cancer, since so many other correlated factors (urban living, smog, tensions, etc.) may also affect cancer.

In general, if factors *A* and *B* fluctuate together, it may be that (1) *A* causes *B*, to be sure, but it might also be that (2) *B* causes *A*, (3) *A* and *B* influence each other continuously or intermittently, (4) *A* and *B* are both caused by *C*, or (5) the correlation is due to chance.

Oversimplification

A common error arises from oversimplifying a subject by omitting essential qualifications. The facts presented may be true in themselves, but if other pertinent facts are omitted, the reader may be misled. Many examples may be found in the pocket-size "quickie" type of magazine

that abounds in fragmentary half-truths. A universal cure is claimed for a certain disease. The article does not mention that its sweeping statements are based on the inconclusive results of experiments on a fragmentary sample of patients or that only partial cures were reported or cures only in mild cases. An insurance company advertises low insurance rates without mentioning that these rates will double after five years. A household freezer is advertised as "featuring exclusive Amanamatic freezing, 2½ times faster!" Faster than what? Another advertisement says, "Dodge Sales Are Up 293% in the Bay Area." Since when?

A former President of the United States, in supporting a wage increase in the steel industry, cited the high profits of this industry without mentioning that these profits were quoted *before taxes*. Taxes actually took away two thirds of these "profits," so that only the remaining third was available for payment of wages.⁷ Another former President announced that unemployment had declined from March to April 1960. This appears favorable, but he neglected to mention that the amount was less than the usual seasonal decline between these months. It is excellent practice to state one's conclusions in simple, nontechnical terms, but not at the expense of overlooking essential limitations and qualifications.

Spurious Accuracy

"There were 90,356,748 motor vehicles registered in the United States during 1965." "The New York Stock Exchange reports 956,804,533 shares traded in 1966 through June 14." "The thirteen regional Shippers Advisory Boards estimated yesterday that railroad freight loadings . . . in the current quarter would be 8,146,723 cars." "A State Industrial Commission study found that a bachelor girl can live a 'single, healthy, and moral' life on a minimum of \$2,422.59 a year." (If she fails to receive that last \$2.59, does her health or morals suffer, or both?) Certainly, none of these figures is correct to the last digit. Such detailed figures are tiresome and suggest a degree of accuracy in counting or measurement that does not exist by any means. The accuracy of economic data is discussed in Chapter 2, where it is suggested that such data in general be rounded off to three or four significant figures.

Assumption of Stability in a Changing Economy

In forward planning, businessmen frequently assume that the most probable future level of activity will be that of the recent past. This is a

⁷ National City Bank of New York, *Monthly Letter on Economic Conditions*, May 1952, pp. 53-56.

fallacy; the normal condition is one of change. For example, an investor buys a bond with the implicit expectation that the purchasing power of the bond will remain relatively stable during its life. If the probabilities point to an inflationary trend in prices, however, as during a war period, he makes a costly mistake in that the proceeds of his bond at maturity will probably buy fewer goods than the same number of dollars would at the time of his investment.

Again, business executives tend to project the current stage of the business cycle into the future. If prosperity exists today, it is assumed to continue tomorrow. Depression today makes men cautious about future commitments. Yet past experience shows that prosperity is frequently followed by "recession," and vice versa.

These examples illustrate the need for forecasting. One of the basic purposes of statistical analysis is to provide a factual basis for planning future operations. Even a crude forecast is likely to be superior to the assumption that past conditions will continue. Applications of statistical analysis in forecasting will be emphasized in this book.

Errors in Percentages

Ratios and percentages seem quite simple, but they are frequently miscalculated through using the wrong base, failing to subtract 100 percent in figuring increases, or misunderstanding the nature of the comparison. A textbook in office management states that "window envelopes cost around \$1.00 less than regular envelopes, or \$3.25, which represents a saving of 76.5 percent." This should be 23.5 percent—or 24 percent to avoid spurious accuracy. A life insurance company reports a gain of insurance in force from \$177 million to \$1 billion in 11 years, or "a gain of 565 percent." This should be 465 percent.

The Ways and Means Committee of the House of Representatives in 1951 considered raising personal income tax rates 3 percentage points "across the board." The tax scale, then graduated from 20 percent up to 91 percent, would be made to run from 23 to 94 percent. Some critics attacked this as a "soak-the-poor measure," since a 3-point increase on the poor man's 20 percent represented a 15 percent jump, while 3 points on the rich man's 91 percent was a mere nudge of 3.3 percent. But other critics claimed that this was a "soak-the-rich" measure, since the poor man's take-home pay would be reduced from 80 to 77 cents on his dollar of income, or only $3\frac{3}{4}$ percent, while the rich man's take-home pay would be cut from 9 to 6 cents, or $33\frac{1}{3}$ percent! The committee compromised by increasing taxes $12\frac{1}{2}$ percent across

the board. This expedient increased the minimum rate from 20 to 22½ percent reasonably enough, but unfortunately boosted the maximum rate from 91 to 102.4 percent! It was subsequently cut to 94½ percent.⁸ This controversy illustrates the importance of the careful use of percentages. Chapter 4 presents an explanation of this deceptively simple topic.

SUMMARY

Statistical analysis is a scientific method of interpreting quantitative data. It is used to draw general inferences by induction from the behavior of variable data, whereas deductive reasoning applies general laws to specific cases. The statistical or inductive method is most effective in the social and biological sciences; the mechanical or deductive method is used more in the physical sciences. Statistical methods have become more important in recent times because of the growth of large-scale production and trade, the increasing scope of government, and improvements in statistical techniques themselves.

Statistical analysis is used in all branches of larger business organizations as a tool of planning and control. The principal statistical activities in business include general business analysis, marketing research, production control, budgeting, personnel and investment studies, credit analysis, and operations research.

Statistical analysis is also widely used in economics and social science generally, particularly in the study of economic fluctuations, social accounting, prices, finance, public utilities, regional analyses, and related subjects. The growth of government activities, too, has required more and better statistics for central planning and administrative purposes.

The basic steps in statistical analysis include (1) *collecting* the data from available sources or sample surveys; (2) *presenting* the results in tables and charts; (3) *analyzing* and *interpreting* the figures by means of statistical techniques, and using the results in (4) *making decisions*, with the aid of probabilities and economic costs or profits. These steps will be followed in this book in the usual order of a statistical investigation.

The true meaning of facts is easily distorted. The statistical investigator therefore must be on guard to avoid misrepresenting the facts and to detect misuses of statistics by others. A critical attitude is essential. The principal pitfalls in the use of statistics are bias, either conscious or unconscious; faulty generalization due to reliance on too small a sample or on one that is not typical of the whole population; faulty deduction

⁸ National City Bank of New York, *Monthly Letter on Economic Conditions*, June 1951, pp. 66-67.

in applying a generalization to exceptional cases; comparisons of non-comparable data; semantic errors such as the use of leading questions; the uncritical inference that correlation between two factors means that one is the cause of the other; oversimplification due to omission of essential qualifications; spurious accuracy; the assumption of future stability in a dynamic economy; and the misuse of ratios and percentages.

PROBLEMS

1. *a)* Explain the meaning of the term "statistics" when used in the singular sense as opposed to its use in the plural sense.
b) Why does the employment of statistical methods in the solution of business problems belong almost exclusively to the twentieth century?
c) Describe the principal statistical activities of a typical large and progressive firm.
2. Locate in the library and give the names of three major statistical journals together with the associations that publish them, and briefly describe the type of subject matter contained therein.
3. Visit an economic research agency or one of the eight types of statistical departments in a business organization mentioned in the text and hand in a two- to three-page outline of its statistical activities.
4. What are some of the principal uses of statistics in economics?
5. Hand in a clipping illustrating an improper use of statistical data. Add a paragraph explaining the type of error presented.
6. Give an illustration of each of the following: (*a*) bias, (*b*) faulty deduction, (*c*) assuming causation from correlation, and (*d*) oversimplification.

SELECTED READINGS

- FERBER, ROBERT, AND VERDOON, P. J. *Research Methods in Economics and Business*. New York: Macmillan, 1962.
 Provides a broad perspective on means of solving research problems.
- GOLDE, ROGER A. *Thinking with Figures in Business*. Reading, Mass.: Addison-Wesley, 1966.
 A primer on "techniques for improving your number sense."
- HUFF, DARRELL. *How to Lie with Statistics*. New York: W. W. Norton, 1954.
 An amusing compendium of statistical misuses.
- KENDALL, M. G., AND BUCKLAND, W. R. *A Dictionary of Statistical Terms*. 2d ed. New York: Hafner, 1957, with *Supplement*, 1960.
 A comprehensive glossary, in English, French, German, Italian, and Spanish.

NEISWANGER, WILLIAM A. *Elementary Statistical Methods*. Rev. ed. New York: Macmillan, 1956.

Chapter 2 contains some excellent illustrations of errors in the use and interpretation of statistics.

REICHMANN, W. J. *Use and Abuse of Statistics*. New York: Oxford University Press, 1962.

A nonmathematical introduction to the meaning of statistical measures.

RIGBY, PAUL H. *Conceptual Foundations of Business Research*. New York: John Wiley, 1965.

Describes the functions of scientific business research as providing the techniques for problem-solving and decision-making, as well as developing new concepts, testing hypotheses, and building models.

ROBERTS, HARRY V. "The New Business Statistics," *Journal of Business of the University of Chicago*, January 1960, pp. 21-30.

Outlines the development of the decision-theory orientation of statistics.

SIELAFF, THEODORE J. *Statistics in Action*. San Jose, California: Lansford Press, 1963.

Twenty-five articles by different authors show how the tools of statistics are used in dealing with business and economic problems.

2. COLLECTION OF DATA

THE FIRST step in statistical analysis is to find the necessary facts. Perhaps they are available in some published source or they may be obtained from the internal records of a business firm. Again, the facts may not be available anywhere but must be collected firsthand in a special survey. For example, one may be asked: "Is the cost of living higher in Chicago than in New York?" "What is the rate of inventory turnover of our copper wire?" "Would sales of our breakfast food be increased by redesigning the package?" The data needed to answer these questions may be obtained from published sources, internal records, and special surveys, respectively.

In nearly every investigation, fact-finding is the first step, and may indeed be the most difficult one. It is nevertheless frequently the most fruitful kind of research. Hence, it is important to know where to find the facts and how to compile them. This chapter describes how to seek out the existing data needed for economic analysis and how to collect, edit, and tabulate original data.

USE OF RESEARCH SOURCES

A thorough search of any previous work that may have been done on a problem is essential as background before a new investigation is undertaken. This requires a careful search of library files. Such available studies may not only provide useful facts but also suggest effective techniques of analysis.

Efficiency in collecting data from libraries comes from learning what data to expect in different sources. While the beginner has no choice but to use what might be called the "shotgun" method—that is, to search until the desired data happen to be found—a seasoned investigator, using a process of elimination based on his previous experience, narrows his

search to two or three likely sources. By contrast, this might be called the "rifle" method. The analyst will then require very little time to find the data, obtain a clue to their location, or discover that they are not available at all.

Ten Steps in Finding a Good Source

There is a general sequence of steps which can be followed in searching for a desired set of business or economic data. The process is one of successive elimination, but some guidance in the order of procedure will facilitate the work. The following procedure is recommended:

Step 1. Consult one or more standard reference sources, such as the *Statistical Abstract of the United States*, *The World Almanac*, the *Information Please Almanac*, and the National Industrial Conference Board's *Economic Almanac*. These are all published annually.

Among the most useful of the monthly sources are the *Survey of Current Business*, *Federal Reserve Bulletin*, *Monthly Labor Review*, *Dun's Review and Modern Industry*, Standard and Poor's *Current Statistics*, and the *Canadian Statistical Review*. See also the monthly bulletins of certain private banks such as the First National City Bank of New York and the Cleveland Trust Company. Charts of leading indexes are published each month in *Economic Indicators* and the *Federal Reserve Chart Book: Financial and Business Statistics*. Regional data are covered in the monthly bulletins of the 12 Federal Reserve banks, the regional commercial banks, and university bureaus of business research. A large number of internationally comparable statistics can be obtained from the *Statistical Yearbook of the United Nations*.

Some of these sources provide statistical supplements. Thus, the *Survey of Current Business* issues a *Weekly Supplement*, a biennial base-book called *Business Statistics*, and supplements entitled *U.S. Income and Output* and *Personal Income by States*. The *Statistical Abstract* publishes *Historical Statistics of the United States* and *County and City Data Book*.

The censuses of population, housing, manufactures, business, transportation, mineral industries, and agriculture provide extremely detailed data on the economy and serve as "bench marks" to check the reliability of the incomplete annual or monthly data which are gathered between censuses.

Step 2. If the data are not found, study the titles, headnotes, footnotes, and references of tables on the general subject to discover original sources which may contain more detail. In turn, study these detailed sources for references to collateral sources.

Step 3. If Steps 1 and 2 have not led directly to a publication containing the information required, consult a bibliography of source material, such as *Selected Business Reference Sources* or *Statistical and Review Issues of Trade and Business Periodicals*, published by the Baker Library of the Harvard Business School. Specialized bibliographies can be found by reference to the *Bibliographic Index* or to special business libraries such as those of the Newark and Cleveland public libraries.

Step 4. Check particularly the source books of federal government statistics. These are described in the Bureau of the Budget's *Statistical Services of the United States Government*. Other government publications may be found in the *Monthly Catalog of United States Government Publications*, *United States Department of Commerce Publications*, the Bureau of the Census *Catalog*, and its *Guide to Industrial Statistics*, all published by the Superintendent of Documents. Andriot's *Guide to U.S. Government Statistics* is useful because it classifies these sources by subject and issuing agency.

Step 5. If the data cannot be located, look up the subject of your inquiry in the library card catalog. Government publications may be listed not under the main subject classification but under "United States" instead. Sublistings are by departments, bureaus, commissions, and offices.

Step 6. If the data are still elusive, or perhaps incomplete, go through the periodical indexes in the library. The following are ordinarily available: *Bulletin of the Public Affairs Information Service*, *Business Periodicals Index*, *The New York Times Index*, and *The Wall Street Journal Index*.

Step 7. Look through trade, financial, and technical magazines related to the subject. Leading weekly publications include the *Commercial and Financial Chronicle*, *Barron's*, *Business Week*, and the London *Economist*. Check the statistical yearbooks and the review numbers of these journals.

Daily papers such as *The Wall Street Journal*, *The Journal of Commerce*, and the financial section of *The New York Times* are invaluable in providing a great variety of up-to-the-minute information. Yet the very promptness of these publications tends to reduce accuracy; hence, the data found in daily papers should be verified, if possible, in other sources before final use.

Step 8. If access to the stacks of the library is possible, go to the section in which you have already found books dealing with the subject. Other publications in the same shelves may contain the desired data.

Step 9. If at this point the desired data have not been found, it is time to consult some specialist who may have knowledge of them. This person may be a research analyst, special librarian, corporation official, trade association secretary, or a government economist in a regional office. A number of such people may be called quickly by telephone. Even though the respondent may not have the desired information himself, he can usually refer the caller to the proper source.

Step 10. Finally, if the desired data cannot be found in published sources, it may be necessary to search for unpublished material from government or nongovernment agencies. In particular, much information can be found directly in the internal records of the company concerned. Many business concerns maintain information centers or specialized libraries that provide a valuable source of data for research projects.

Unpublished data for many industries can be secured from the Department of Commerce or from the various trade associations. A leading source of older unpublished records of the federal government is the National Archives in Washington, D.C. For historical studies in economics, business, sociology, and political science, and for detailed data on the two world wars, this storehouse of records is especially valuable.

Only in the most difficult cases will it be necessary to employ all of the foregoing steps. Usually the first two or three will be productive. After a few searches have been made, the general contents of the major publications will be sufficiently familiar so that in most cases the proper source can be selected immediately.

Checking for Discrepancies

Once the data are collected, they should be examined to detect discrepancies and then verified by cross reference when several sources are available. Discrepancies in data may appear as a result of changes in unit, in coverage, or definition; revisions; or typographical errors.

Changes in the nature of the *unit* may ruin a series. Thus, Table 831 of the 1965 *Statistical Abstract*, showing the number of four-engine aircraft in commercial service from 1955 to date, has little significance in itself because of the great improvement in aircraft performance between the era of the DC-4 and that of the jet DC-8 or Boeing 707. The data must be used in conjunction with figures on available seats, speed, mileage, and the like.

Discrepancies due to changes in the *coverage* or scope of a series may be illustrated by the widening of boundaries of cities, which introduces

errors into many kinds of urban data. For example, census reports of the increase in the population of many cities between 1960 and 1966 have been exaggerated because these cities meanwhile annexed outlying areas, whose population was included in the 1966 figures but not in the 1960 figures. Another change in scope occurred when the F. W. Dodge Corporation reports of construction contract awards were expanded in coverage from 37 to 48 states beginning in 1957.

Changes in *definition* appear in the *Census of Manufactures*, which includes a reclassification of establishments by industry in accordance with the *Standard Industrial Classification Manual*, a redefinition of the minimum size limit for establishments included, and other changes in concept.

Revisions. An example of discrepancies due to revisions in data is found in the comprehensive overhaul of gross national product estimates published in the August 1965 *Survey of Current Business*. Thus, the component "changes in business inventories" for 1964 was changed from \$3.7 to \$4.8 billion, an increase of 30 percent. When the newspapers report a change of 1 or 2 percent in gross national product accounts (or most other business indicators) as being significant, therefore, it must be remembered that such a change might easily be due to errors of estimate. Since many statistics are first released in preliminary form and later revised as more returns are received, the latest available figures, of course, should be used.

Typographical Errors. Typographical errors occur in every newspaper, journal, or book, almost without exception. Arithmetic errors are also common. They may best be discovered by checking the figures with other data and by examining each statement critically, rather than accepting the results without benefit of mental evaluation.

Cross Reference. Frequently, similar data are collected by several agencies. In these instances, the sources should be compared to determine which is most complete, which contains the data in most usable form, and which has the best general record of reliability. Also, if the sources show different figures, they may reveal the types of discrepancies enumerated above.

As an example of the use of cross reference, suppose you wished to compare the volume of passenger traffic by rail, bus, and air in 1953.¹ You find the figures reported by three leading trade associations, as shown in Table 2-1.

These figures are quite different. The largest figure for bus travel, for

¹ This example is taken from *Management Methods*, December 1955, through courtesy of Management Magazines, Inc.

example, exceeds the smallest by 44 percent. Yet all are issued by reputable organizations and are based on reports of the Interstate Commerce Commission and other government agencies.

Some of the discrepancies are due to a simple matter of timing. In July 1955 the ICC issued revised figures for 1953, taking into account certain types of travel which they had previously ignored. The AAR issued a mimeographed sheet in which they reported the revised series, but the other two associations had just released the annual editions of their statistical fact books, not so easily changed, and they were still issuing figures based on the older data.

Table 2-1
PASSENGER TRAFFIC BY RAIL, BUS, AND AIR—1953
(Billions of Passenger-Miles)

	SOURCE OF DATA		
	Association of American Railroads	National Association of Motor Bus Operators	Air Transport Association
Rail.....	32.7	27.5	26.9
Bus.....	28.4	21.3	19.7
Air.....	17.4	15.6	14.7
Total.....	78.5	64.4	61.3

But this does not explain the entire difference. The ATA did not include travel on the nonscheduled airlines, since their statisticians did not feel the figures were very reliable. *Bus Facts*, the NAMBO publication, included the travel of rail *commuters*, which the others did not, and it also added an estimate of chartered buses.

In other words, no two of the associations were really talking about the same thing. In general terms, of course, they were all discussing passenger travel, but each had its own special definition, which was just different enough to throw the figures off substantially.

In summary, then, the rule is to use the latest reliable source which is available and whenever possible verify it by cross reference, carefully investigating any discrepancies that cannot readily be explained.

Judging the Accuracy of Economic Data

Economic data vary widely in their accuracy, even though they may appear to be exact. Thus, Sears, Roebuck and Co. on January 31, 1966, reported total assets of \$4,909,324,502; but with all the difficulties of

evaluating securities, real estate, and other accounts, only about four figures—4,909 millions—could be considered meaningful in appraising the company's balance sheet. The rounding error of 1 part in 10,000 is surely negligible. In fact, most economic data should be rounded off to three or four significant figures for simplicity in tabulation, computation, and interpretation.² Additional figures are neither valid nor necessary in decision-making (though they may be needed for accounting consistency).

On the other hand, many reported figures are subject to much wider errors than three or four significant figures would indicate. Therefore, it is important to estimate the size and type of errors inherent in the basic data. This may be done by studying the nature of the original data, the collection process, and the purpose for which the figures were gathered. For example, the *Survey of Current Business* in February 1967 reported that the value of new construction put in place in January 1967 amounted to \$4,630 million. This might appear to be an exact figure, but actually it represents estimates by more than a dozen collection agencies derived from hundreds of different sources of varying reliability. "Construction takes place on widely scattered sites and is carried on by tens of thousands of small contractors and by persons doing their own building and repair work,"³ so that the above figure may be considerably in error. In order to understand the nature and limitations of basic statistics, therefore, one should study the text and footnotes accompanying a report, check other sources, and write the original collection agency, if necessary, for a description of its methods.

Sometimes the errors in data are estimated by the collection agency itself. For example, in "Consumer Income in 1964," the Bureau of the Census says: "Since the estimates in this report are based on a sample, they . . . are subject to errors of response and nonreporting and to sampling variability."⁴ This is followed by a discussion of errors of response, and a table and explanation of the "Standard Error of Estimated Percentage" (explained in Chapter 13) as a measure of sampling variability.

² The following rules are recommended for rounding numbers: (a) When a number greater than five is dropped, increase the preceding digit by one. (b) When a number less than five is dropped, leave the preceding digit unchanged. (c) When the exact number five is dropped, increase the preceding digit by one if it is an odd number but leave it unchanged if it is an even number. That is, the rounded number is always even. This rule prevents cumulative errors in addition.

³ F. C. Mills and C. D. Long, *The Statistical Agencies of the Federal Government* (New York: National Bureau of Economic Research, 1949), p. 60 and Chart 3.

⁴ *Current Population Reports*, Series P-60, No. 47, September 24, 1965, p. 21.

The U.S. Bureau of Labor Statistics, too, warns that unemployment figures for small subgroups of the population for one month are unreliable. Yet when it reported that Negro unemployment had risen from 8.4 percent in June 1965 to 9.1 percent in July, at the time of the Watts riots in Los Angeles, a number of writers cited these figures to prove that the expansion of the economy had left the Negro behind. Later though, the August figure was reported as 7.6 percent, and subsequent months were even lower. The July figure was a statistical blooper.

It is an excellent rule for the business analyst, therefore, to estimate the error in any figures he prepares or uses, so that he may avoid being misled by unreliable data.

Significant Figures in Computation

Two rules should be observed in performing basic calculations with approximate numbers:

1. In addition or subtraction, the result should contain no more decimal places than the least accurate of the numbers themselves. Thus, the *World Almanac* reported the area of Europe as 3,769,107 square miles, and that of Asia as 17,300,000 square miles (i.e., estimated to the nearest 100,000). The total for Eurasia should then be stated as 21,100,000, not 21,069,107, square miles.

When applied to subtraction, however, this rule reveals a pitfall: A relatively small error in two large figures may produce a large percentage error in the difference. To illustrate, the number of unemployed persons in the nation is sometimes estimated by subtracting the number employed from the total labor force of those available for jobs. Suppose employment and labor force are each subject to an error of one million, or $1\frac{1}{4}$ percent, in either direction. Then the resulting estimate of unemployment may be off two million, or 100 percent, as shown below.

<i>Estimates of</i>	<i>Millions of Persons</i>	<i>Possible Error (Percent)</i>
Labor force.....	80 ± 1	$1\frac{1}{4}$
Employment.....	78 ± 1	$1\frac{1}{4}$
Unemployment.....	2 ± 2	100

This simple arithmetic accounts for the wide errors that frequently occur in estimates of unemployment, the federal deficit, personal savings, net profits of corporations, and similar values obtained by subtraction.

2. In multiplication and division, the result has no more significant figures than the *least* number of significant figures in the numbers themselves. (Significant figures are the digits that show the extent to which a number is accurate, excluding the zeros used to fix the position of the decimal point.) As an example, *Economic Indicators* reports net farm income at \$12.1 billion in 1965, with an estimated 3.5 million farms, so that net income per farm (the quotient) is \$3,457. However, if the number of farms is significant to only two figures, then only the first two figures in net income per farm are significant. This is because the number 3.5 represents any value between 3.45 and 3.55. Dividing these end values into \$12.1 billion income gives a range of from \$3,507 to \$3,408 in net income per farm. These possible values differ even in the second significant figure.

Squares and square roots, as special cases of multiplication and division, should contain no more significant figures than the original number. Thus, $(26.8)^2 = 718$, and $\sqrt{26.8} = 5.18$.

In more extended calculations, however, the figures should not be rounded off until the final result is stated. This is to avoid cumulating the errors of rounding in subsequent operations of multiplication or subtraction.

COLLECTION OF ORIGINAL DATA

We have described how to find and use existing data in research sources. In case the figures are not already available, however, they may have to be collected directly by a survey of the original source. This section describes how to plan and carry out such a survey.

Most surveys are concerned with human populations. General Motors polls the public to determine its likes and dislikes in car styling. The J. Walter Thompson Company maintains a consumer panel of selected families to check the brands of food products being purchased. Market research generally utilizes consumer surveys to measure the market acceptance of a product. Public opinion polls cover every possible topic. "The questionnaire is to our civilization what art and philosophy were to the Greeks, or law and sewers to the Romans—a natural form of self-expression."⁵

Many surveys, however, relate to nonhuman populations. The quality control supervisor of a manufacturing company samples its products to check for defective items. The purchasing agent does the same for goods being bought. The auditor samples a "population" of inventory items to

⁵ Dwight Macdonald in *The New Yorker*, November 22, 1958, p. 89.

check average costs. This section is primarily concerned with the collection of data from people, but the principles discussed apply to the collection of other types of original data as well.

Census versus Sample

Some investigations require a complete enumeration or "census." The U.S. *Census of Population*, for instance, is a complete enumeration. Other complete collections of data, such as the statistics of incomes, cigarette consumption, and gasoline consumption, are by-products of the tax-collecting function of the government.

In contrast to these complete censuses are the great majority of surveys which depend upon obtaining a sample which will be typical of the whole population. For example, the Bureau of the Census estimates the number of cars and other durable goods that American consumers plan to buy during the coming year from a sample of only 17,000 households out of the 53 million households in the country—only 1/30 of 1 percent of the total.⁶ Similarly, the U.S. Department of Agriculture uses a sample of two quarts of grain in a carload (57,600 quarts) to determine the grade of the grain; and the U.S. Bureau of Labor Statistics Consumer Price Index is based on prices of a few hundred commodities and services obtained from a relatively small number of stores and other respondents.

There are three basic reasons for the widespread use of sampling:

1. Sampling usually saves a great deal of time and money. Often, when the cost of a complete census would be prohibitive, the necessary information can be obtained from a sample. The results of a survey need only be accurate enough to provide an adequate basis for decision-making. Beyond a certain point the increase in information from additional data is not worth the increase in cost.
2. In many cases, a complete census is impossible as, for example, in making a quick check of consumer preferences for an entirely new product, or in the destructive testing required to determine the breaking strength of steel rods, or in measuring the effectiveness of a new antibiotic.
3. Finally, sampling may actually yield more accurate results than a complete survey. A small group of interviewers can be selected and trained more rigorously to reduce the biases in a survey than a very large staff. Similarly, in testing materials, a few careful measurements may be preferable to a larger number of crude measurements. Improvements in

⁶ *Federal Reserve Bulletin*, September 1960, pp. 977-1003.

sampling techniques, too, have led to many advances in modern survey methods.

Personal Interviews versus Mail Questionnaires

Original data are usually collected either through personal interviews or through questionnaires sent out by mail. These methods are compared below.

Personal Interviews. The principal advantage of personal interviews lies in the opportunity to secure nearly complete returns from the desired sample. Interviewers can usually reach nearly all of the people selected as a typical sample of the population to be surveyed.

When mail questionnaires are used, on the other hand, a large proportion of the recipients may disregard them. Thus, there is no assurance that those who reply are typical of the entire group to whom the questionnaires were mailed. Frequently, those who cannot give a favorable response will not reply at all. Or, those with more education are more likely to reply than others that one wishes to reach. Finally, questionnaires may be answered by a business subordinate or a junior member of a family rather than by the person to whom they are addressed. This situation creates an error quite apart from any tendency of respondents to give biased answers, a difficulty which the investigator faces in any case.

In the second place, personal interviewers can generally obtain accurate replies through explaining the questions, persuading the informant to provide the desired information, and judging the validity of the response. If the respondent appears uninformed or facetious, for example, the interviewer can discount his reply. Of course, the interviewers themselves must be carefully selected and trained to avoid introducing their own biases in phrasing the questions or recording the answers.

The advantages of personal contact are lost in the mail questionnaire. Not only will many questionnaires be discarded, but a number of those that are returned will have been misinterpreted or only partially completed, particularly if the list of questions is a long one.

Interviewing may also be done by telephone. This method makes it possible to obtain a large number of interviews quickly and at a relatively low cost. However, it is limited to telephone subscribers, who may not be typical of the entire population. Furthermore, only a relatively small amount of information can be obtained in each call. It is also difficult to obtain such data as age, economic condition, or occupation over the telephone.

Mail Questionnaires. The principal advantage of obtaining information by mail is, of course, its economy. The cost of mailing, including return postage, is only a few cents per questionnaire, so that even if only a few replies are received, the cost per return will generally be less than that of personal interviews. Hence, this method is used whenever the results are believed to be reliable.

Mail questionnaires are particularly economical if a large geographical area is to be covered. While interviews may be employed economically within a single locality, their use may be too costly if extensive travel is required.

The use of mail questionnaires may also be preferable to that of interviewers in case the respondent requires considerable time to compile the data, as in reporting the operating results of retail stores. Interviewers ordinarily can only collect data that are immediately available, while questionnaires can be answered at the respondent's convenience.

In large surveys that would require numerous interviewers who cannot be thoroughly trained, a mail questionnaire has an advantage in avoiding the interviewers' bias. This was a factor in the Census Bureau's decision to use mail questionnaires extensively in place of interviews for the 1960 Census of Population.

Sometimes a "consumer panel" of typical families is selected by personal interview, and then these families are induced to report their brand purchases monthly by mail. The inducement may be money, merchandise, or stamps exchangeable for goods. This method combines the economy of mailing methods with the accuracy of a personally selected sample.

Mailing lists for questionnaires can be obtained from city directories, telephone books, Dun & Bradstreet's credit rating books, trade directories, city and county records such as lists of taxpayers, automobile registrations and building permits, the membership rolls of various organizations, and commercial mailing-list dealers. Of course, any such list must be checked to be sure that it is accurate, complete, and up to date.

Sometimes a questionnaire is sent to an entire mailing list; then, later on, interviewers are sent out to visit a number of the persons who did not reply. In this way it is possible to determine whether the replies of the nonrespondents differ from those of the respondents and, if so, in what respects. This combination method minimizes costs without sacrificing too much reliability. A variation of this method is to have per-

sonal interviewers collect data in thickly populated centers and to send questionnaires by mail to respondents in less accessible areas.

The variety of methods used in collecting data may be illustrated by the four major services that rate TV shows by size of audience, according to an article in *TV Guide*. The American Research Bureau sends "diaries" to some 2,200 homes; the families record their viewing and mail in reports weekly. *Pulse* sends interviewers (mostly women who are local residents) to people's homes; about 150,000 interviews are conducted monthly. *Trendex* uses the spot-check technique of telephone calls to some 1,000 TV homes in 15 cities. A. C. Nielsen Company has installed about 1,200 "Audimeters" on TV sets in homes scattered throughout the country. The audimeter records on tape the channel to which the set is tuned and the time it is turned on and off.

Preparation of Questionnaires

The success of a survey depends to a large extent upon the quality of the questions used. The type of question included will depend upon whether interviewers or mail questionnaires are employed. Interviewers can generally obtain replies to questions which are more involved and more personal than those on mail questionnaires. In spite of this difference, the two types of questions can best be discussed together with separate explanations as needed.

A common practice is to test a preliminary draft of questions by submitting them to a small test group of persons similar to those in the sample selected. Such a "pretest" will aid in revising the questions and in improving the interview technique. Specifically, the pretest should check whether these seven rules have been followed successfully: (1) organize the questions carefully, (2) use clear wording, (3) define terms, (4) be brief, (5) avoid offensive questions, (6) avoid bias, and (7) provide adequate instructions. These rules are discussed below.

1. **Organize the Questions Carefully.** In outlining the content of a questionnaire, include only those questions which contribute directly to the objective of the survey. Begin with questions that identify and describe the respondent and then list the major information questions. Defer personal or controversial opinions to the end of the list. Two different questions may be included on the same subject to provide a cross check on important points.

2. **Use Clear Wording.** Each question should contain but one idea. It must be stated as simply as possible, so that there can be no doubt in the mind of the respondent what is wanted. For example, the 1950 Census of Population included a question on employment for per-

sons 14 years old and over. Yet, the simple query, "Are you employed or unemployed?" could not be asked because it is ambiguous. Housewives, college students, temporarily unemployed persons, or those on leave from their work might classify themselves in either category. Again, the numbers of unemployed persons who are looking for work, are unable to work, or are retired have very different significance. The question was therefore phrased specifically as follows (in the follow-up questionnaire used for persons not at home when the agent called) :

What were you doing last week? (Check each box that applies to you.)

- a) ☐ I worked at a job or in my business or profession or on a farm.
- b) ☐ I was looking for work.
- c) ☐ I had a job, profession, or business from which I was temporarily absent.
- d) ☐ I did housework in my own home.
- e) ☐ I am permanently unable to work.
- f) ☐ None of the above applies to me.

When interviewers are used, the questions may be abbreviated, since the interviewers are already familiar with the meaning of each question and the definition of terms. On the other hand, a mail questionnaire must be filled out by the respondent himself, so the questions must be complete sentences, as above, and must make their own appeal.

3. *Define Terms.* In preparing a questionnaire, any word, phrase, or unit of numerical data must be so precisely defined that no ambiguity exists and no technical uses of terms are unexplained.

Some units of numerical data have standard definitions, such as the dollar or the short ton. Others must be defined specifically wherever they are used. Thus, a room in a dwelling is not a usable unit until many borderline cases such as breakfast nooks and utility rooms have been either included or excluded by definition.

4. *Be Brief.* The use of a few easily answered questions in a questionnaire will increase the number of replies. Questions should be worded so that they can be answered by "yes" or "no," by numbers, or by a simple check to record choices, provided such answers are adequate. Requests for historical information should be avoided if possible, since this is often difficult to recall. Make sure that no repetition of information is requested except in the case of "check" questions.

5. *Avoid Offensive Questions.* Great care must be taken to avoid offense. For example, in small, closely held businesses, one cannot ask the question, "What was the dollar value of your net sales last year?" But the approximate data may be obtained by asking, "Please indicate in which of the broad groups below your net sales for last year would fall,"

followed by several sales classes arranged to give enough detail for use in the subsequent analysis. Questions concerning personal income, morality, or religion should be avoided, if possible.

6. *Avoid Bias.* Bias may enter in two ways. First, the question may be phrased so as to suggest a certain answer. An example of a biased wording is "Did the frozen peas taste better to you than canned peas or dried peas?" This is the notorious "leading question." It would be much better to list the three types of prepared peas and request that the user number them in the order of preference.

Second, estimates that are based on opinions rather than on actual figures may be biased. Suppose you were inquiring of a manufacturer of drugs whether his product was distributed at retail mainly through chain stores or independent stores. His direct contacts with the buyers of chain retailers might lead him to suppose that they were his chief customers, whereas a study of the sales records might well show the reverse. Questions should be objective rather than subjective.

Respondents may have unconscious biases about their own attitudes or actions. For this reason it is sometimes better to use indirect questions to obtain information. Thus, in a survey of consumer preferences, it was found that the question "What do you think *your neighbor* would like in his next automobile (chrome, space, economy, etc.)?" produced more unbiased replies than "What would *you* like in your next automobile?"

7. *Provide Adequate Instructions.* The instructions for interviewers must contain not only all definitions of terms and all fixed procedure for interviewing but also a description of cases to be included, boundaries of areas, time scheduling, and other pertinent details. Interviewers should be chosen for their ability to inspire confidence and to obtain information without offense. Those selected must be given rigorous training in all phases of the survey, including field training if necessary.

A mail questionnaire should be accompanied by a letter of transmittal containing a brief explanation of the purpose of the survey and providing some incentive for answering it, such as (1) an appeal for cooperation, (2) mutual interest in the results, (3) possible profit from the results, (4) obligation to the investigator, (5) prestige of position held by respondent, or (6) a gift of merchandise, stamps, or funds.

Follow-up Procedure

In some surveys, response to the first attempt to collect the information, whether by mail or personal interview, is insufficient for the purpose of the investigation and a follow-up procedure is needed. The extent of the follow-up is determined by the rate of response to the

initial inquiry, the costs involved, and the precision required in the final results. Follow-ups may be conducted by a method that differs from that of the original survey, as in the case where persons failing to reply to a mail questionnaire are interviewed in person.

Editing Schedules

As the returns from a survey come in, they must be edited very carefully in order to detect any irregularities in the responses. The editor should search out and correct any omissions or inconsistencies in the returns, verify check questions and computations, and make sure that all questions have been interpreted in a uniform manner.

After the various irregularities have been adjusted, the editor must reclassify the items, if necessary, in the form in which they are to be tabulated. Sometimes the process of preparing the returns for tabulation involves *coding*, or assigning numbers to different replies. Coding of information to be transferred to punch cards or magnetic tape is a necessary step in mechanical tabulation, which is described in the next section.

Preliminary Tabulation

The next step is to transfer the edited information to preliminary tables. Out of these detailed tables come the final tables for analysis and presentation. The three principal methods of transferring information from the collection forms to preliminary tables are (1) hand tabulation, (2) punch cards, and (3) electronic data processing.

Hand Tabulation. Returns can be tabulated by sorting and counting individual cards or by using tally sheets. The sorting-counting process can be used to advantage when the raw data are relatively simple so that each case can be recorded on one card. The cards can be sorted and subsorted into piles according to any desired plan of classifying the data. The number of cards in each pile can then be recorded on a suitable form.

The use of tally sheets differs from sorting-counting in that the schedule cards or sheets are not separated into piles according to the various classifications. Instead, blank forms are made up to conform to the classifications of the data. The information is then tallied on the form as it is read from the questionnaire.

Punch Cards. When a great many questionnaires are to be analyzed, machine tabulation may be necessary. Equipment is available to perform quickly and accurately the steps of sorting, counting, cross tabulating, and recording in columnar form. These advantages have led

many business concerns to install punch-card systems for maintaining records of current operations, issuing bills to customers, or mailing dividend checks.

The basic principle of punch-card tabulation is that a hole punched in a card represents, by its horizontal and vertical position, a certain statistical fact. It becomes a permanent record that can be tabulated at any time by running the card through a machine.

Chart 2-1 represents an example of a punch card used in an industrial market survey.

Chart 2-1
STANDARD PUNCH CARD

[illegible]

The information for this card is collected by a salesman of industrial equipment to provide the customer information needed in estimating future sales potentials. The customers are other companies that buy this equipment.

The 80 columns of the card are grouped into fields of information, as shown in the headings. A code must then be set up to transfer descriptive information to the card. Thus, the salesman's branch office is coded "707," as shown by the rectangular punches in the first three columns. Numerical data, however, can be punched directly—for example, 31,500 employees in columns 35 to 37. The completed card shows the branch office, sales representative, customer, his state, county, industry and department of the company, the number of employees (to indicate the size of the company), and sales of the manufacturer's equipment lines A and B at various periods.

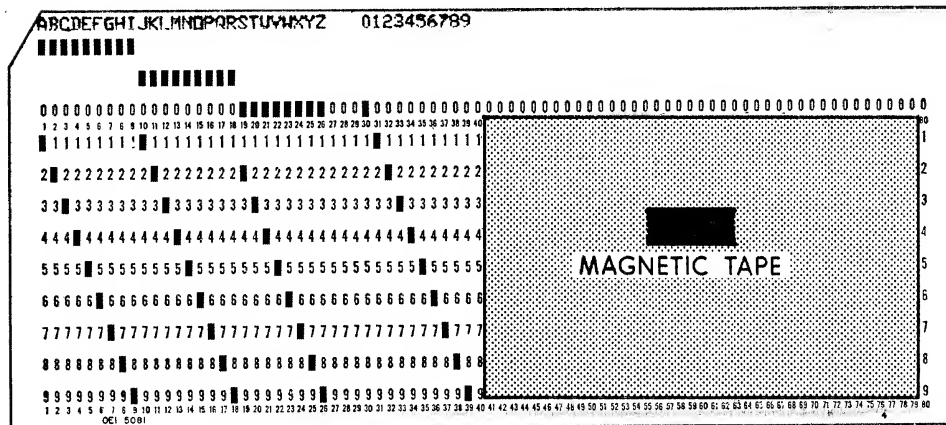
After similar cards have been punched for hundreds of customers, the cards can be quickly sorted and tabulated to show present and "poten-

tial" sales classified by branch office, by sales representative, by customer, by lines (A and B), by industry, by state and county, or on whatever basis the data are needed.

Electronic Data Processing. Electronic machines now perform high-speed calculations that obviate the need for sorting and tabulating equipment in many statistical operations. Electronic equipment can do everything the older equipment can do at much higher speeds with greatly expanded capacities and, in addition, can perform the most intricate calculations.

Chart 2-2

RELATIVE SIZE OF STANDARD PUNCH CARD AND MAGNETIC TAPE
WITH EQUIVALENT CAPACITY (SCALE REDUCED)



All necessary statistical data and instructions as to what to do with them are fed into these new machines, usually on magnetized tape or punch cards. Magnetized tape has much greater capacity and is, therefore, increasingly preferred to the use of the punch cards for this purpose. Chart 2-2 illustrates the difference in capacity: a small fraction of an inch of tape can carry the same information as a standard punch card. The machines then sort, classify, and tabulate data, or perform series of calculations in a fraction of a second each, by means of electronic impulses transmitted through intricate systems of transistors or tubes which can store or "remember" numbers and use them in successive operations.

A general-purpose electronic computer can perform a wide variety of functions in data processing: It can sort the information as desired, convert it into a different form, store it for future use, transfer it to other locations in the system, perform all types of arithmetic computations, and print the final results in readable form. All of this is done at high speeds in a completely integrated operation, with no human interven-

tion. Furthermore, the machine can make simple decisions, as in reviewing payroll time sheets to determine whether employees are eligible for overtime pay. The versatility and speed of electronic data processing systems are therefore revolutionizing large-scale data handling and decision-making in modern business.

A detailed description of electronic data processing and computing is beyond the scope of this book. For a good nontechnical discussion of these systems from the management viewpoint, consult the Chapin or Gregory and Van Horn references at the end of this chapter. See also the monthly journals *Data Processing Digest* and *Automation*.

The business executive does not need to become an expert in electronics or mathematics in order to use electronic data processing. He can deal with the problems involved by acquainting himself with the general capabilities and limitations of these machines. For the actual programming he can rely on technical experts.

SUMMARY

A knowledge of research sources is essential in business or economic analysis. The first step in using these sources is to find the necessary materials. This may be done by consulting any of several types of sources, such as statistical reference books, bibliographies, card catalogs, periodical indexes, trade journals, library stacks, and experts in the field.

In collecting data from published sources, great care must be taken to test the figures for accuracy and for validity by noting any changes in the units in which the data are expressed, shifts in coverage, revisions, or typographical errors. It is particularly useful to check several publications against each other and to study the method of collecting the data as a means of detecting errors and estimating the reliability of the results.

The accuracy of figures must always be considered. Economic data are seldom accurate to more than three or four significant figures, so longer numbers should ordinarily be rounded off. The accuracy of any figure can be estimated by studying the method of collection.

The number of significant figures in computations is governed by the minimum number of significant figures in the data being processed. In subtraction, however, small errors in the original figures may produce a much larger error in the difference.

If the necessary figures cannot be found in published sources or in the internal records of a business, a special survey must be made. Such a survey need not be a complete census but can be restricted to a limited

group if the respondents represent a typical sample of the entire population under study.

If personal interviewers are used, they can canvass the entire group to be sampled; they can explain questions carefully and evaluate the replies, thereby securing more reliable results than is possible by mail questionnaires. On the other hand, mail questionnaires are generally more economical, particularly if a wide area must be covered; so they are ordinarily used if the results can be made reliable. A combination of these two methods is sometimes used. Occasionally, too, interviews may be conducted by telephone.

In preparing questionnaires, it is essential to organize the questions carefully, to avoid ambiguities in the wording, to define all terms and units used, to avoid offensiveness and bias, to provide adequate instructions, and still be as brief as possible.

After the questionnaires have been filled in and returned, they must be edited for irregularities and prepared in proper form for tabulation.

The data compiled in simple projects can be tabulated by entering the necessary information on cards and sorting them by hand or listing and totaling the figures on large tally sheets. For more complex investigations, the data can be coded and entered on punch cards. These cards are punched, checked, sorted, tabulated, and totaled by special machines. Finally, electronic data processing machines have been developed in recent years for the high-speed tabulation and computation of complex data.

PROBLEMS

1. A young stockbroker interested in general business conditions is planning a small library of statistical source material. The following list has been selected as adequate: *Economic Almanac*, current year; subscriptions to *Monthly Labor Review*, *Economic Indicators*, and *The Wall Street Journal*; *Moody's Manuals*, most recent volumes; and *Census of Business*.
 - a) Which of the foregoing would you retain?
 - b) Name four others that should be included.
 - c) Give reasons for your choice in (a) and (b).
 2. Name the publications that correspond to the following descriptions:
 - a) Published monthly by a banking agency in Washington and containing some text material and detailed tables that are practically identical in form from month to month, chiefly on the subject of finance.
 - b) A monthly publication of the congressional Joint Economic Committee which includes charts and tables on prices, employment, production, national income, purchasing power, and finance.
 - c) An annual issue of a monthly magazine giving yearly estimates of per-
-

- sonal income and retail and wholesale sales for all counties in the United States. Useful in marketing studies.
- d) An annual volume containing general tables that show yearly quantity and value of mineral production by states; also employment and injuries. Separate chapters on each mineral give domestic production and shipments, prices, consumption, stocks, foreign trade, and world production by country.
 - e) A series of volumes which present detailed data for 1954, 1958, and 1963 on manufacturing activities in the United States.
3. Name a source in which you think each of the following sets of data would be available. Explain your choice in each case.
- a) The number of tons of primary aluminum produced in the United States monthly during the past year, including the latest month available.
 - b) The latest data on the number of employees on the payrolls of manufacturing concerns, by industries, in the United States.
 - c) The amount of sales by apparel stores in the state of New York in 1963.
 - d) The number of freight carloadings of livestock shipped in the United States during the last year.
 - e) The index of industrial production in the United States for the most recent month.
4. The answer to each of the following questions is to be found in a commonly used *government* source. Give exact reference to the source.
- a) The percentage of increase in population for New York and for Texas from 1950 to 1960.
 - b) An index of consumer prices in Chicago during the most recent month and the same month last year.
 - c) The wholesale price per gallon of No. 2 fuel oil at New York Harbor for the most recent week.
5. The answer to each of the following questions is to be found in a commonly used *nongovernment* source. Give exact reference to the source.
- a) The number of new passenger car registrations for Ford and Chevrolet last year.
 - b) The number of business failures in manufacturing in the United States during the latest month available.
 - c) The percentage of foreign-made trucks sold in the United States each year since 1954.
6. Certain difficulties of collection occur in each of the following problems. Find as much information as you can in answering the question and explain the circumstances in the sources that make it difficult to secure complete and comparable data.
- a) Compare the number of savings banks, depositors, and amount of savings in your own state with another state as of a recent date.
 - b) Compare the changes in the number of employees in the chemical industry and in the automobile industry over the past 50 years.
-

- c) Compare the number of full-time employees in one-, two-, and three-store independent food stores in the United States with the number employed in chain food stores for selected years since 1954.
- d) Select the five industries whose indexes of employment were lowest during the most recent month and compare these figures with their indexes in 1932.
7. Round each of the following numbers to (a) four significant figures and (b) three significant figures:
- | | |
|-------------|--------------|
| (1) 395.890 | (5) 547,550 |
| (2) 5,064.1 | (6) 6,274.78 |
| (3) 75.682 | (7) 594,681 |
| (4) 10,072 | (8) 87.463 |
8. How many figures would you expect to be accurate in each of the following? Give reasons for your answer in each case. (All examples were taken from the *Statistical Abstract of the United States, 1965*.)
- a) The population of the United States was enumerated on April 1, 1960, as 179,323,175 persons.
- b) The population of the United States on April 1, 1965, was estimated by the Bureau of the Census as 194,032,000 persons.
- c) The Office of Education reports the enrollment in colleges, universities, and professional schools in 1962 as 3,726,000.
- d) The total assets of all member banks of the Federal Reserve System on March 31, 1965, were \$285,300,000,000.
- e) The Department of Commerce estimates from a sample survey that the total retail sales of the United States in 1964 amounted to \$261,630,000,000.
9. Find the value of a wheat crop estimated at 3,500 bushels at a probable price of $\$2.16\frac{7}{8}$ per bushel. Express the result to the correct number of significant figures.
10. For the year ended January 31, 1965, Sears, Roebuck and Co. reported income before federal income taxes of \$551,243,707, less provision for federal income taxes of \$247,150,000, equals net income of \$304,093,707, or \$4.00 per share of stock. Express to the correct number of significant figures: (a) the net income and (b) the estimated number of shares outstanding.
11. State in each of the following examples of collection whether personal interviews or mail questionnaires are preferable and whether the census or sample method should be used. Give reasons for answers in each case.
- a) A retail dry goods association wished to study the distribution of operating expenses of its 61 members.
- b) A marketing research agency wished to inquire from the owners of a certain make of refrigerator whether they would purchase the same make again.
-

- c) A corporation president wanted information concerning how many of its 15,400 employees were homeowners, the value of their homes, the amount of mortgage, the interest rate paid, and the monthly payment on the mortgage.
12. What is the purpose of pretesting a questionnaire before starting a mail survey?
 13. Cite the three most important rules, in your opinion, that should be followed in preparing a questionnaire on consumer attitudes toward color television. Give reasons for your choice.
 14. Explain which of the following alternative wordings is preferable for a questionnaire and why:
 - a) (1) What body style do you prefer for your next automobile?
 - (2) Check the body style you prefer for your next automobile:

4-door sedan_____	Station wagon_____
2-door sedan_____	Convertible_____
Hard top_____	Other (specify)_____
 - b) (1) Do any of the following apply to your concern? (Check which.)

Clerks poorly trained_____
Clerical overtime pay too high_____
Too many clerks_____
Office management inefficient_____
 - (2) Which of the following would be most effective in reducing office expenses in your concern? (Check one.)

Additional training of clerical employees_____
Reduction of paid overtime for clerks_____
Reduction of clerical office force_____
Reorganization of office force_____
 15. Define the following terms for use in a questionnaire. Be sure to provide for possible borderline cases: (a) a household, (b) a wholesaler, (c) an unemployed person, and (d) a drugstore.
 16. The following card was returned to the interviewer by the editor. What do you think the editor found wrong, and what did he want the interviewer to do?

RESIDENTIAL VACANCY SURVEY	
Address <u>324 Henkel Circle front and rear houses</u>	Serial No. _____
Ward _____ Tract _____ Enumeration District _____	
No. of Dwelling Places in Building	
One _____ Two _____ Three _____	
Four _____ Over Four (give number) <u>5</u>	
Occupied <u>3</u>	Vacant <u>1</u>
Residential <u>X</u>	Combination <u>X</u>
Agent _____	<u>R. A. Shawn</u>

17. Which of the three methods of preliminary tabulation (hand tabulation, punch cards, or electronic data processing) would you use for each of the three surveys described in Problem 11 above? Explain your choice in each case. Assume the following number of returns: 11 (a) 52 dealers; 11 (b) 5,120 refrigerator owners; and 11 (c) 1,200 employees.
18. Visit a nearby electronic data processing installation and prepare a brief report on (a) description of the computer, (b) the types of operations performed, and (c) savings in costs, as compared with earlier methods.

SELECTED READINGS

BROWN, LYNDON O. *Marketing and Distribution Research*. 3d ed. New York: Ronald Press, 1955.

Describes how to plan a survey, prepare a questionnaire, and edit and tabulate the results.

CHAPIN, NED. *An Introduction to Automatic Computers*. 2d ed. Princeton, New Jersey: Van Nostrand, 1963.

Discusses computers from a systems point of view, with emphasis on data processing uses.

DAVIS, GORDON B. *An Introduction to Electronic Computers*. New York: McGraw-Hill, 1965.

Covers the basic features and concepts of a number of computer systems, with particular reference to business problems.

GREGORY, R. H., AND VAN HORN, R. L. *Business Data Processing and Programming*. Belmont, California: Wadsworth Publishing, 1963.

A compact, introductory book on data processors and their programming for business.

MORGENSTERN, OSKAR. *On the Accuracy of Economic Observations*. 2d ed. Princeton, New Jersey: Princeton University Press, 1963.

A penetrating analysis of the many inaccuracies of economic statistics.

STOCKTON, JOHN R. *Business Statistics*. 3d ed. Cincinnati: South-Western Publishing, 1966.

Chapter 2 presents a detailed account of problems in assembling and tabulating data.

WASSERMAN, PAUL, ET AL. *Statistics Sources*. 2d ed. Detroit: Gale Research, 1965.

Lists over 9,000 sources of statistics, with dates and publishers' addresses, in the United States and abroad.

3. EFFECTIVE USE OF TABLES AND CHARTS

THE COLLECTION of data has been covered in the previous chapter. This chapter describes the methods of preparing data for *analysis* and *presentation* in the form of tables and charts. The facts of business must be tabulated and charted properly before they can be clearly interpreted.

The first problem of presentation is: Should data be presented in the form of a table or a chart?

Tables have several advantages over charts: (1) more information can be presented, (2) exact values can be read from a table, and (3) less work is involved in preparation. On the other hand, charts have the advantages of (1) attracting attention more readily with a graphic picture and (2) showing trends and comparisons more vividly than the abstract figures in tables. Most readers are visual-minded and prefer graphs to figures.

In many reports, a chart and a table are placed together so that the reader can see both the general picture and the detailed figures.

STATISTICAL TABLES

A statistical table is a classification of related numerical facts in vertical columns and horizontal rows. Classification is the grouping of facts into classes that are distinguished by some significant characteristic. The classes should be defined so as to be mutually exclusive; hence an item cannot be included in two or more classes.

Methods of Classifying Data

The three common bases of classification are qualitative differences, size, and time. They are illustrated in Table 3-1, which compares unemployment rates by sex, age, and race for 1963, 1964, and 1965.

Classification based on *qualitative* differences is illustrated by the breakdowns by sex and race. The distinction is one of kind rather than of amount. Other qualitative classifications could be made by marital status or occupation. Geographical classifications are also qualitative. Thus, unemployment rates could be reported by states or by metropolitan areas. The use of ratios, rates, and percentages to compare qualitative characteristics is considered in Chapter 4.

Table 3-1
UNEMPLOYMENT RATES IN THE UNITED STATES, 1963-65
(As Percent of Labor Force)

	1963	1964	1965
Total.....	5.7	5.2	4.6
Male.....	5.3	4.7	4.0
14 to 19 years of age.....	15.5	14.5	13.1
20 and over.....	4.5	3.9	3.2
White.....	4.7	4.2	3.6
Nonwhite.....	10.6	9.1	7.6
Female.....	6.5	6.2	5.5
14 to 19 years of age.....	15.7	15.0	14.3
20 and over.....	5.4	5.2	4.5
White.....	5.8	5.5	5.0
Nonwhite.....	11.3	10.8	9.3

SOURCE: *Survey of Current Business*, January 1966.

The breakdown of unemployment rates by age groups illustrates classification by *size* or magnitude. Similarly, the unemployed could be classified by years of education or by number of weeks out of work. Size classifications will be analyzed in Chapters 4, 5, and 6 by means of frequency distributions, averages, and dispersion measures.

The columns showing 1963, 1964, and 1965 represent a *time* classification or time series. Time series may be further divided into (a) measurements taken at different points of time, like population, prices, or the data in Table 3-1 and (b) cumulative data that build up from zero in a given period, like monthly steel production or weekly retail sales. Methods specially designed for studying time series are presented in Chapters 18 to 21.

Reference and Summary Tables

There are two principal types of tables, depending on the purposes for which they will be used. These are reference and summary tables.

Reference tables are sometimes called general-purpose tables or repository tables. They are designed to present information for general use, without applying it to any particular problem. Such tables are to be found in the *Statistical Abstract of the United States*, the various censuses, and other government publications. Reference tables are frequently detailed so as to provide complete data for a variety of purposes, and they may include definitions, description of the collection process, and other information. They are not intended to be read through but are arranged for easy reference to the information they contain. Such tables are commonly found in the appendixes of business reports; their use in the body of a report should be avoided because they may be unduly cumbersome.

Summary tables are sometimes called special-purpose, derived, or text tables. They are designed to present specific figures for some particular use. These tables are usually short and appear in the body of a report to illustrate some point in the text. The tables in this chapter are of this type. Summary tables should be simple and attractive in form, to hold the reader's attention. They must be arranged so as to emphasize the most important figures presented and to point up significant comparisons.

A summary table is often abstracted from one or more reference tables. In the process of preparing a summary table from a reference table, it is often desirable to (1) select only the important figures, (2) use group totals instead of detailed data, (3) round off all numbers to three or four significant figures, (4) rearrange the data to place the most important item at the top left for emphasis, (5) place related figures next to each other for easy comparison, and (6) provide ratios, averages, or totals to aid in summarizing and interpreting the results. This trimming and rearranging will add greatly to the effectiveness of any summary table.

Construction of Tables

The following principles of construction have proved useful in making effective tables. The table should have unity. To avoid confusing the reader with diversive ideas, all entries should be pertinent to the subject of analysis.

Cross classification is required to facilitate study of various combinations of characteristics and to focus attention on the main comparisons. Simple classification is illustrated by General Motors sales in Table 3-2 (ignoring for a moment the other companies), which are classified according to a single characteristic—body type.

If classification is desired according to two characteristics simulta-

neously, such as body type and manufacturer—they must be *cross classified* in a “two-way” table. In Table 3–2, the manufacturers are listed horizontally in the column heads, while body types appear down the left-hand side in the stub. The other principal parts of a table are also designated in this example.

Three or more orders of classification can be shown by subdividing either or both of the first two classifications. For example, in the stub of Table 3–2, each body type could be subdivided into four-door and

Table 3–2

AUTOMOBILE SALES, BY BODY TYPE
AND MANUFACTURER, 1966
(Thousands of Cars)

←Title

Body Type	General Motors	Ford	Chrysler	Total	←Column Head
Sedans.....	000	000	000	000	←Row
Hard Tops.....	000	000	000	000	
Other Types.....	000	000	000	000	
Total.....	000	000	000	000	

↑
Stub↑
Column

two-door models. However, when the further subdivision of data leads to tables which are too complex to be read easily, it is preferable to increase the number of tables. Do not spend time devising ways of presenting multiple classifications in a single table; make two or more tables instead.

The title of a table should be both simple and complete. For simplicity, a brief catch title or narrative title is sometimes used in the first line, followed by a detailed subtitle.

Ordinarily, a complete title should answer the questions: (1) “*What* do the data represent?” (automobile sales); (2) “*Where* are the data from?” (three American manufacturers); (3) “*When?*” (1966); and (4) “*How classified?*” (by body type and company).

The unit of measure should always be explicitly stated. Thus, “Thousands of Cars” appears under the title of Table 3–2.

Footnotes or headnotes (just below the title) should be used to explain anything in a table that cannot be understood by the reader from the title, column heads, and stub. These notes should contain statements concerning figures that are missing, preliminary, or revised and explanations concerning any unusual features of the table that are not self-evident.

A table should always give exact reference to the sources from which

the data were taken. There are three reasons for this: (1) the reader is given a sound basis for evaluating the data; (2) the reader is able to find further information if needed; and (3) the author gives proper credit to the source and places on it the responsibility for any error in the original data.

The arrangement of a table contributes to its effectiveness. In particular, the data should be arranged so as to emphasize important points. The table should also include interpretive figures such as totals, percentages, and averages. Furthermore, since the space for each entry is wider than it is high, in a cross classification the longer list of items ordinarily appears in the stub. It is also better to use the longer wording in the stub, if possible, to avoid crowding the narrow column headings. In any case, the most important figures to be compared should be placed in adjoining columns or rows.

CHARTS

Charts are designed to serve either of two major purposes: (1) analysis or (2) presentation of data.

As Tools of Analysis

Charts may be used as working tools of analysis in any of the following ways: (1) as the first step in an investigation, the analyst can use a graph as a visual guide in planning the mathematical computations and general procedure of a research study; (2) later the chart provides him with a step-by-step picture of developments, thus aiding him in the use of his judgment and in checking the accuracy of the results; (3) graphic measurement may be used in place of mathematical computation to save time and labor, as in a ratio chart or nomograph; (4) freehand curves may be fitted to data in more varied and flexible forms than mathematical curves, as in trend and regression analysis.

Some types of charts are especially useful as analytic tools. In particular, the graph of a frequency distribution, a time series plotted on a ratio chart, and a scatter diagram of two related variables are all essential in statistical analysis. Graphic methods of analysis will be used throughout this book.

For Presentation of Data

A chart or graph¹ is also a most vivid and forceful medium for the presentation of statistical data. The reader gains a clear and simple

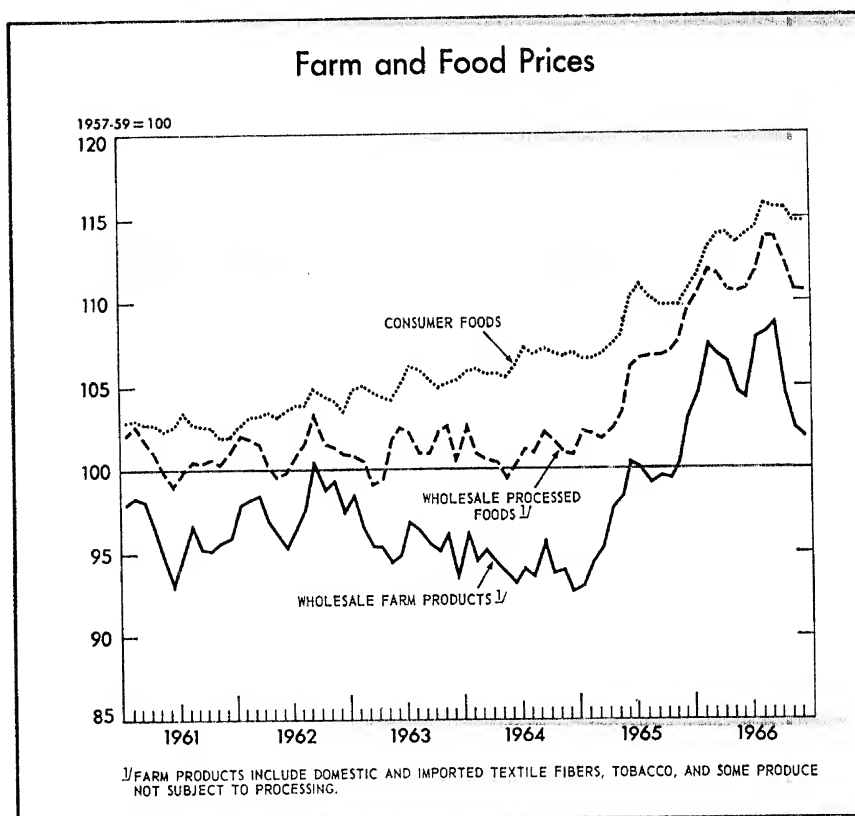
¹ "Graph" may be used in the same sense as "chart" to mean any representation of statistical data in pictorial form or it may refer to a line or curve drawn upon a chart.

impression from a chart which he cannot get from reading the same material in a table or text. With proper planning and execution, a chart will give a truthful, clear, and attractive picture of the facts. A poor chart, on the other hand, may completely nullify the effect of statistical analysis. The following sections show how to construct and interpret the principal types of charts.

Chart 3-1 illustrates the fundamentals of good presentation: the chart is simple in its detail and terminology, accurate in presenting a clear-cut picture with specific labels and scales, large enough in size for easy reading, and properly proportioned. The three curves stand out clearly, and are differentiated for proper emphasis.

It is often desirable to "tell a complete story" graphically by combining a number of charts in a sequence so as to form a connected narrative. A running narrative title is used to tie the charts together in a

Chart 3-1



SOURCE: U.S. Department of Labor, as reported in *Economic Report of the President*, January 1967, p. 88.

complete and unified exposition. This device is used widely in business journals and government reports.

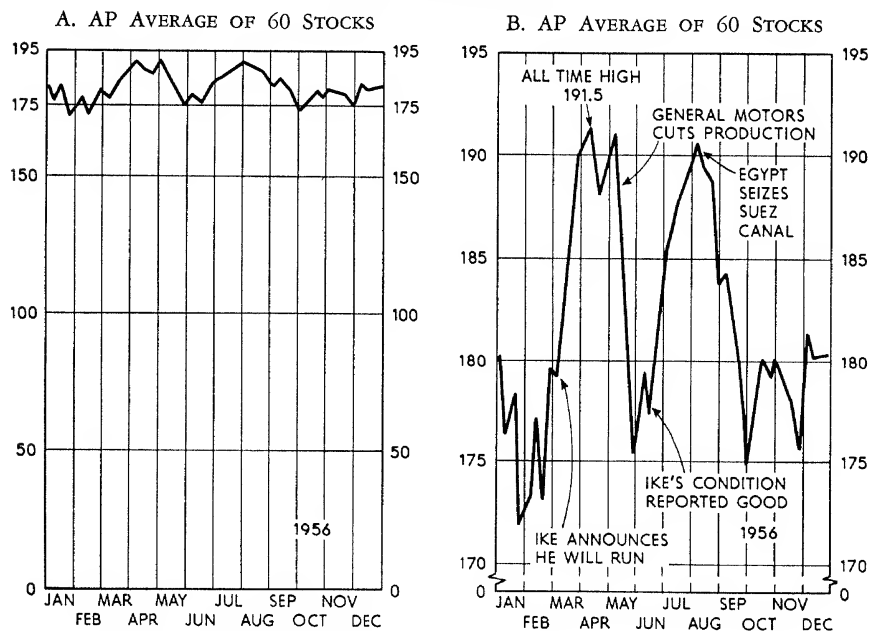
Scale Distortion

It is particularly important that a chart have proper proportions, because scale distortion can twist the meaning of a finished diagram. Suppose we wish to chart the course of common stock prices during 1956. Chart 3-2, panel A, shows the weekly movements of the Associated Press average of 60 stocks, plotted with a complete vertical scale extending from 0 to 195. The market was apparently quite stable; the AP average began and ended the year at 180 and never departed from this level by as much as 7 percent. But this picture is rather tame for a press release. By cutting off the unused vertical scale below 170 and stretching the remaining scale about seven times, the Associated Press draftsman produced the chart shown in panel B. The market now appears to have experienced a succession of soaring booms and precipitous collapses! This is truly spectacular, but is it the truth?

The apparent fluctuations in stock prices may also be affected by the type of average used. Thus, the Dow-Jones average of 30 industrial

Chart 3-2

SCALE DISTORTION



SOURCE: Panel B reproduced from Associated Press release.

stocks has gyrated so widely because of its high level—about \$970 at the end of 1965—that the New York Stock Exchange, on December 31, 1965, instituted its own index of 1,254 common stocks beginning at the \$50-a-share level. The new index, of course, is more stable than the Dow-Jones average in terms of dollar movements because of its lower base.

COMMON TYPES OF CHARTS

Three principal types of charts are used in business and economics: *Line* charts consist of a series of points connected by straight lines. Such a set of connected straight lines is usually referred to as a "curve." The scales may be either arithmetic or logarithmic. In the case of the semilogarithmic or *ratio* chart, only one scale is logarithmic, while the other is arithmetic. *Bar* charts may consist of vertical bars called columns, horizontal bars, or pictorial figures arranged as bars. *Scatter diagrams* consist of dots which show the relationship of two variables in regression analysis.

This section will briefly describe arithmetic line charts and bar charts, since these types are simple and generally well understood. The ratio chart will be discussed more fully, since this is an important analytic tool whose characteristics and uses are often misunderstood. Scatter diagrams will be described in Chapter 22.

ARITHMETIC LINE CHARTS

The line chart with arithmetic scales is by far the most common type in general use. The plotted curve effectively shows the absolute magnitudes and trends, provided the proportions are not distorted.

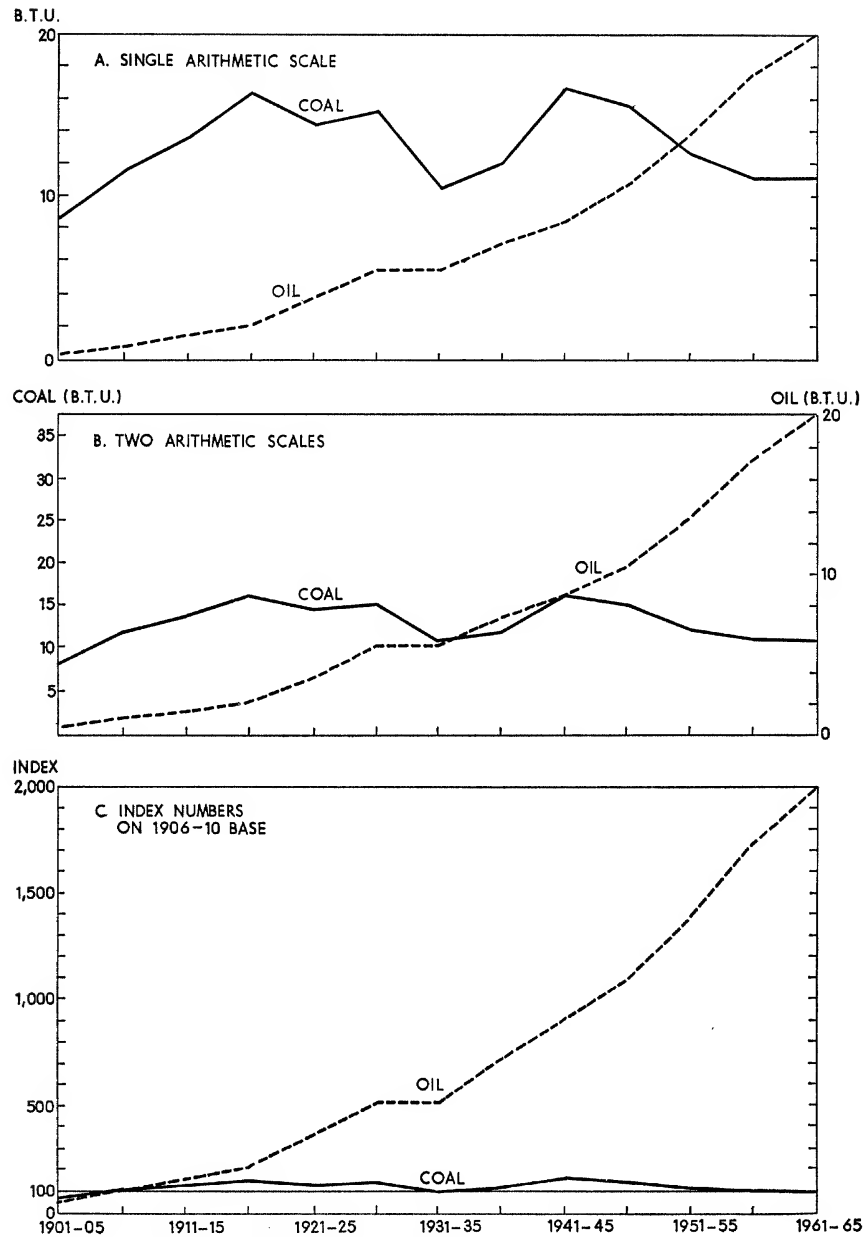
In a time series graph the horizontal scale shows the time units from left to right while the vertical scale measures the amount, as in the first three charts of this chapter. This usage follows the convention that the independent variable should be plotted on the X axis and the dependent variable on the Y axis.

Methods of Comparing Several Series

A problem arises in comparing two or more series recorded either in different units (such as number of employees and dollars of sales) or in the same unit at levels so far apart that it is difficult to use the same scale effectively for both (e.g., total industry sales versus sales of a small company). Chart 3-3 shows three ways of comparing two series on arithmetic scales.

Chart 3-3

SUPPLY OF ENERGY FROM COAL AND DOMESTIC OIL
FIVE-YEAR AVERAGES, 1901-65
(In Quadrillions of British Thermal Units)



SOURCE: *Statistical Abstract of the United States.*

Single Scale. If the two series are in the same unit and are not too far apart in size, a single arithmetic scale is best even though it minimizes the fluctuations of the smaller series. In this case, if the purpose is to show that coal has been the major source of power, as compared with domestic oil, prior to the 1950's, the top scale (Chart 3-3A) is the one to use.

Use of Two Scales. If the series are recorded in different units, or are widely different in size, they may be brought close together for easy comparison by plotting them on different scales. These scales should be selected so that the average level of the two is about the same. Thus, by setting 1.8 Btu's of coal equal to 1 Btu of oil in Chart 3-3B (since the average level of coal output has averaged about 1.8 times that of oil for the period 1901-65), the curves are brought closer together. The timing and general direction of ups and downs can now be compared, but the amounts of change are not at all comparable. Although such charts are sometimes justifiable, this scale adjustment should be avoided if possible, because it may mislead the reader.

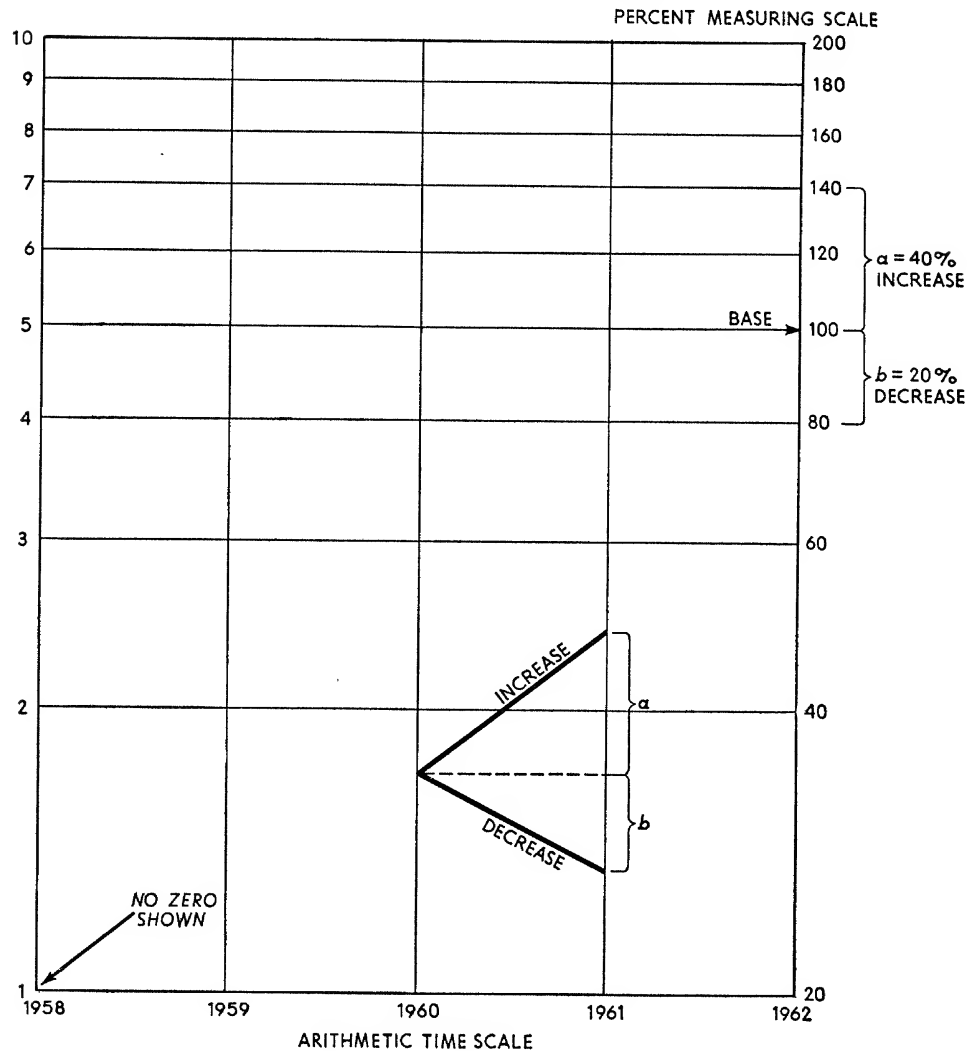
Index Numbers. When a comparison of relative changes of the variables is needed, they may be reduced to index numbers, using the same base period in each case. Indexes are percentages obtained by dividing each series by its value in the base period. The 100 percent or base line and the entire percent scale will be common to the several series. Chart 3-3C compares the percentage variation of the two sources of power relative to the base period 1906-10. An index number chart affords valid comparisons between any period and the base period, but not necessarily with other periods when the indexes may be far apart. Thus, in 1941-45 coal scored a greater gain than oil over the preceding period both in Btu's and in percent, but the index number chart shows oil rising more steeply, because of its higher level relative to the 1906-10 base. Index numbers will be discussed further in Chapter 18.

Ratio Scale. Perhaps the best method of comparing the relative changes in two dissimilar series is to plot them both on a ratio scale. This method permits percentage comparisons between *any* two periods, as described below.

RATIO CHARTS

Although arithmetic scales are satisfactory for showing absolute changes in the data, they fail to reveal clearly what is often of more importance—the relative or percentage changes. For example, it is ordinarily not so significant that a company's sales increased more

Chart 3-4

ONE-CYCLE RATIO OR SEMILOGARITHMIC CHART
WITH PERCENT MEASURING SCALE

dollars over a given period than those of its smaller competitor as that its *percent* increase was greater. For many purposes, then, relative comparisons are more important than absolute comparisons. The ratio chart has come into widespread use for showing relative changes and comparisons, since it is superior to the arithmetic chart in this respect.

The term "ratio chart" means that the chart shows ratios in their true proportion; that is, equal ratios or percents cover equal spaces on the vertical scale. The ratio chart is also called a "semilogarithmic" or

"semilog" chart because the natural numbers are plotted on the vertical scale at distances from the "1" bottom line proportional to their logarithms, while the horizontal axis shows time on the usual arithmetic scale. Thus, in Chart 3-4, the scale number "1" is at the bottom (since $\log 1 = 0$) and the top number 10 is one unit above (since $\log 10 = 1$), the unit being 5 in. in this diagram. The "2" is marked .301 of the way up the graph (since $\log 2 = .301$ in Appendix B), or 1.5 in. up; "3" is marked .477 of the way up; and so on. However, since only natural values are plotted, it is no more necessary to know logarithms in using a ratio chart than in using a slide rule. In fact, the ratio scale on a chart is the same as that on a slide rule.

A ratio chart should be so labeled, but if not, it may be identified in a publication by the fact that the vertical scale numbers get closer together as the scale rises. In particular, the vertical distances between 1 and 2, 3 and 6, and 5 and 10 are all the same, since these distances all represent the same ratio of 1 to 2 irrespective of their position on the chart.

In the ratio chart (as the term is generally used) only one scale is logarithmic. The double logarithmic chart, in which both scales are logarithmic, will be discussed in Chapter 24 in connection with scatter diagrams showing the relationship between two variables. (Many types of logarithmic grids are made by Keuffel and Esser, Dietzgen, Codex, and other manufacturers.)

A log scale is said to have one *cycle* if the scale numbers extend only from 1 to 10 (or multiples thereof); two cycles if the scale is divided into two equal parts covering the ranges 1 to 10 and 10 to 100, respectively; three cycles if divided into three equal parts ranging from 1 to 10, 10 to 100, and 100 to 1,000; and so on. The scale can also be extended downward indefinitely to 0.1, 0.01, 0.001, etc., but can never reach zero. Hence, the log scale cannot be used for a series that includes zero or negative values.

How to Plot

The first problem in plotting data on a ratio chart is to choose between one-, two-, and three-cycle paper. If the largest value in a series is less than ten times the smallest, one-cycle paper is usually preferable, because this has the largest scale. Only the portion of the scale that is used in plotting need be shown in the finished chart, since there is no zero or other base line from which heights are measured.

The printed log scale begins with 1, rather than 0, at the bottom. In order to plot data most easily, mark the bottom line with one of the numbers 1, 2, 4, or 5, followed or preceded by any number of zeros,

such as 0.01 million persons, 20 dollars, 4,000 tons, or 5 percent. If some other value, such as 3 or 75, were placed at the bottom, it would complicate plotting, since the minor grid lines would represent odd amounts. If there is a choice of two numbers, select the one that will best center the curve on the chart.

Once the bottom value is selected—say \$20—multiply this by the printed scale figures 1, 2, 3, . . . and mark them accordingly (20, 40, 60, . . .) until the top of the cycle is marked with a value ten times the bottom (200). This is a *must*. If the printed figures 1, 2, 3, were labeled 20, 30, 40, for example, the logarithmic proportions would be lost and the graph would be meaningless as a ratio chart. Special care must be taken in plotting data because some printed grid lines are omitted as the scale contracts in the higher values.

Different scales can be used to compare series of disparate size or those expressed in different units. For example, the relative growth of a large and a small company, or of coal production in tons and oil in barrels, may be fairly gauged because the slopes of the curves register percentage changes, which are comparable even if the original units are not. Thus, the incompatible are made compatible.

The scales should be selected so as to bring the series close together for easy comparison, with the more important series on top for appearance's sake. The choice of scale affects only the height of a curve above the bottom line, which is not significant; it does not affect the shape of the curve in any way.

Uses of the Ratio Chart

The slope of a line on a ratio chart indicates the percent change between two points of time. A continuing line of the same slope or two parallel lines therefore represent the same relative movement. The steeper the slope, the greater the percent rate of change. A given vertical distance corresponds to the same percent difference anywhere on the chart. These characteristics give ratio charts the unique advantages described below.

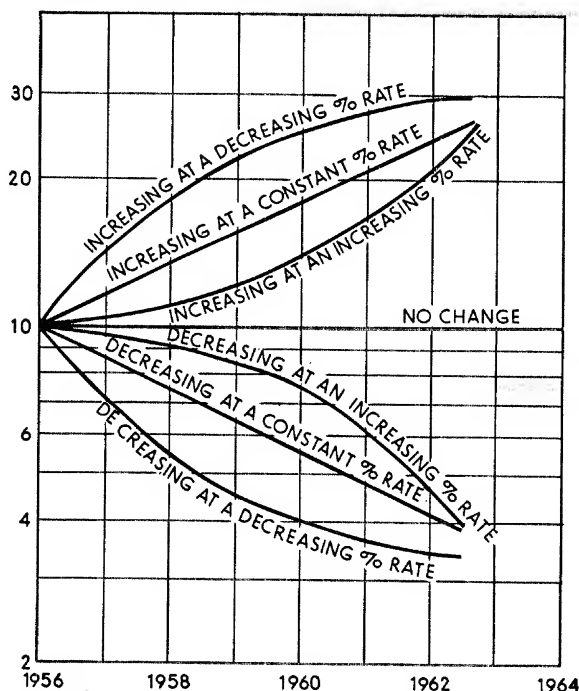
Constant Rate of Growth as a Straight Line. A series growing or declining by the same percent each year, such as a sum of money at compound interest, or sales increasing 10 percent a year, appears on the ratio chart as a straight line.² If the series curves away from the straight

² This "logarithmic straight line," also called an exponential curve or compound interest curve, fits any geometric progression, such as 1, 2, 4, 8, 16. It should not be confused with a line representing a constant *amount* of change, or arithmetic progression, such as 1, 2, 3, 4, 5, which appears as a straight line on an *arithmetic* grid.

line, it denotes a corresponding change in the rate of growth or the rate of decline, as shown in Chart 3-5. Many young industries expand at about a constant percent rate each year until they mature, when the rate of growth tends to taper off, as in the top curve of Chart 3-5. Thus, the oil production curve in Chart 3-6 is nearly straight from 1900 to 1925 but bends over to the right thereafter, while the older coal industry

Chart 3-5

MEANING OF CURVE SHAPES ON RATIO CHART



grew at a decreasing rate until about 1920, and then it turned down.

By watching a company's production curve on a ratio chart, therefore, the analyst can determine whether or not it is maintaining its past rate of gain. Furthermore, if historic factors of growth may be expected to persist, the analyst can project past trends in order to forecast future output. This method is described in Chapter 19.

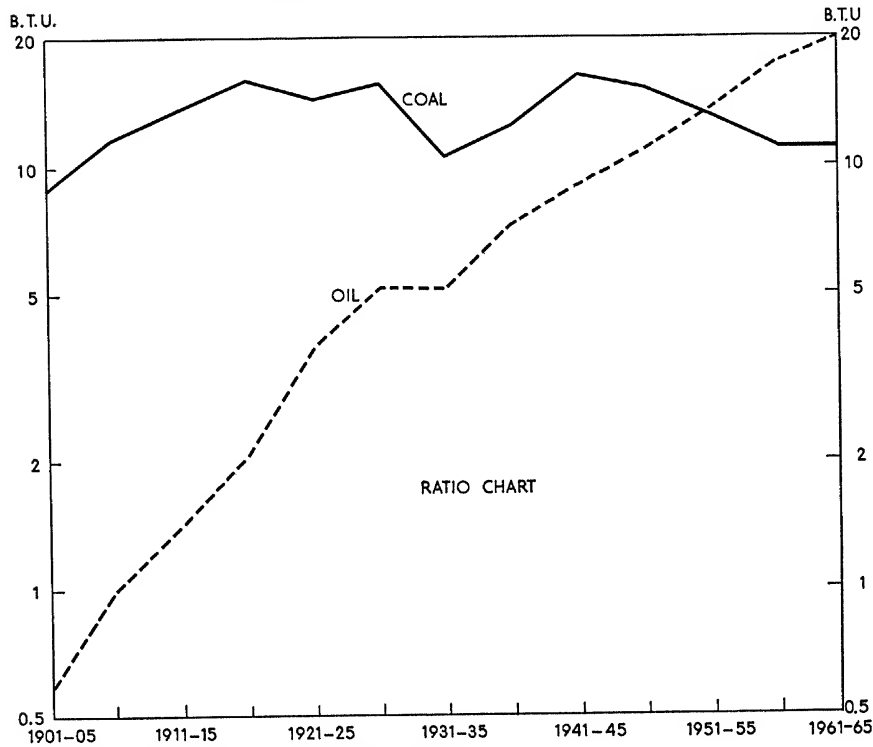
Comparison between Two Curves. The relative growth or decline of two or more curves can be seen at a glance by comparing their slopes on a ratio chart. Thus, in Chart 3-6, if the oil production curve rises more steeply than the coal production curve, this means that its percent growth is greater, irrespective of the size of the two series or the

units in which they are measured. Oil has gained steadily at the expense of coal since 1900 except during World War II, when oil was rationed because of war needs.

An arithmetic graph of two series on a single scale always emphasizes

Chart 3-6

SUPPLY OF ENERGY FROM COAL AND DOMESTIC OIL
FIVE-YEAR AVERAGES, 1901-65
(In Quadrillions of British Thermal Units)



SOURCE: *Statistical Abstract of the United States.*

the growth of the larger one, as in Chart 3-3A. Or, if two different scales are used to bring the curves together, the relationship is arbitrarily distorted (Chart 3-3B). Even index numbers only afford easy comparison with one base level (Chart 3-3C); if some earlier period had been taken as base, the relative increase in the use of oil would have been much greater because of the smaller base. The ratio chart affords true relative comparisons between any two points on the grid, and yet absolute values can be read from the scale, unlike the case of index numbers.

Performing Calculations on a Ratio Chart. Percentages or ratios may be read directly from a log scale in this way:

1. Mark a percent measuring scale as shown on the right column of Chart 3-4. That is, on a one-cycle chart, multiply the printed scale numbers by 20, so that the scale extends from 20 percent to 200 percent. On a two-cycle chart, mark the center line "100 percent," and so on, so that the vertical scale extends from 10 to 1,000 percent.

2. Mark the *vertical* distance between any two points on the edge of a blank strip of paper, or take it off on a pair of dividers (e.g., the increase *a* or decrease *b* between 1960 and 1961 on Chart 3-4.

3. Lay off the increase upward, or the decrease downward, *from the 100 percent base point* of the measuring scale, and read the value of the second point as a percent in terms of the first point as 100 percent. The percent *change* is this figure minus 100. Thus, on Chart 3-4, the 1960-61 increase *a* is read off as 40 percent, while the decrease *b* is 20 percent.

Instead of transferring the vertical distances on a chart to its own percent measuring scale, a separate strip of the graph paper marked with a percent scale may be placed vertically on the chart to measure percents or ratios directly.

Limitations of Ratio Charts

Ratio charts have certain limitations in the presentation of data which restrict their use accordingly: (1) They do not give a visual idea of absolute magnitude as a distance above the base line, although these magnitudes can be read from the scale. (2) They are difficult for the layman to understand and so should not be used for simple illustrations which an arithmetic chart could show as well. (3) They cannot show zero or negative values. (4) Finally, they are sometimes mistakenly used to contract a wide range of absolute values into a small space. This is legitimate if relative movements are of interest, but if a picture of absolute changes is needed, an arithmetic scale should be used.

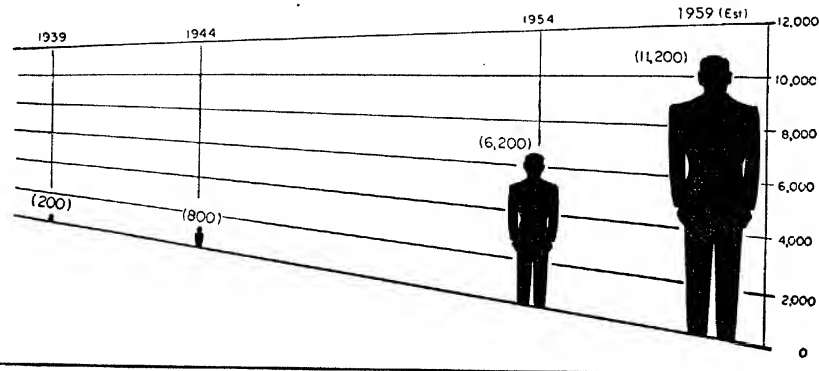
BAR CHARTS

While the arithmetic line chart is the most important type for the presentation of data, various geometric forms are also in common use for popular portrayal of simple comparisons. These forms may be of one dimension, such as bars of uniform width which *vary only in length*; two dimensions, such as circles; or three dimensions, such as spheres, or the human figures shown in Chart 3-7.

Of these forms, bars usually give the most accurate impression of size,

since in two- or three-dimensional figures the reader is uncertain whether to compare the diameters or the areas or the volumes, as the case may be. If the diameters denote the true comparison, then the areas or volumes exaggerate it. Chart 3-7, for example, was planned to indicate that the need for "avionics" engineers (i.e., those who develop control systems for aircraft) was expected to nearly double in a five-year period. The right-hand silhouette was therefore drawn about twice the height of the adjoining one. However, the engineer of the future is

Chart 3-7

Airframe firms' employment of avionics engineers

SOURCE: *Aviation Week*, December 27, 1954, p. 42.

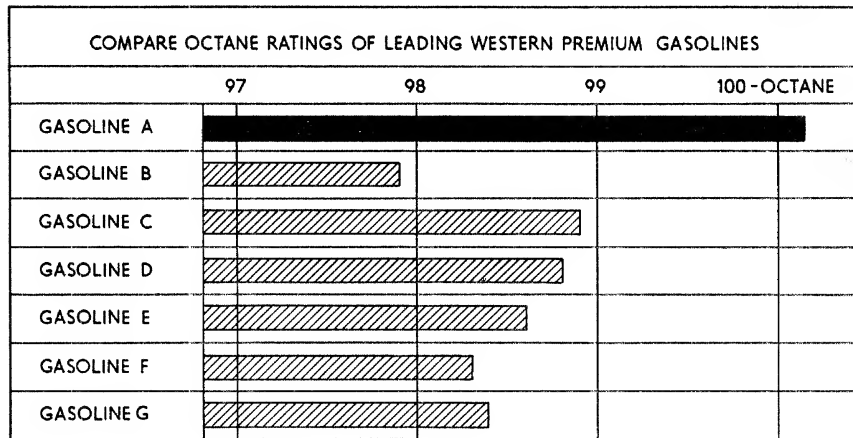
shown as *four* times as big as his predecessor in area and nearly *eight* times as big in weight—since this is a three-dimensional measure. The drawing thus greatly exaggerated the increase in the demand for engineers. (The increase was still further exaggerated by drawing the chart in perspective.) For this reason the use of two- and three-dimensional drawings of different size should generally be avoided. If a pictorial diagram is desired, it is preferable to show rows of figures of uniform size, depicting engineers or whatever is appropriate, since the length of the row indicates the amount in the same way as a bar does.

Bar charts may be preferable to line charts in portraying a relatively few values of one or two series. Line charts are preferable where there are many values or several series. Bars emphasize the individual amounts, while lines emphasize the general trend. Bars are also effective for showing the component parts of a whole. Bars and lines may be combined, as in a time series, where bars may represent yearly averages for earlier years and a curve shows the more recent monthly movements.

The bars are usually *vertical* in time series (Chart 3-9) and in frequency distributions (Chapter 4). They are usually *horizontal* in qualitative comparisons (Chart 3-8). Percent changes are also represented by horizontal bars extending from a vertical base line to the right, if positive, or to the left, if negative, and arrayed in order of size.

Since bars represent magnitudes by their lengths, the zero line must be shown and the arithmetic scale must not be broken, in order to

Chart 3-8



present a true comparison.³ The bars would be shortened by the same amount, to be sure, but their proportional difference would be increased. Note the scale in Chart 3-8, which is taken from a newspaper advertisement for "Gasoline A." By omitting the scale values from 0 to nearly 97, the octane rating of Gasoline A is made to appear 200 percent greater than that of Gasoline B—about 100 times the true difference of 2 percent!

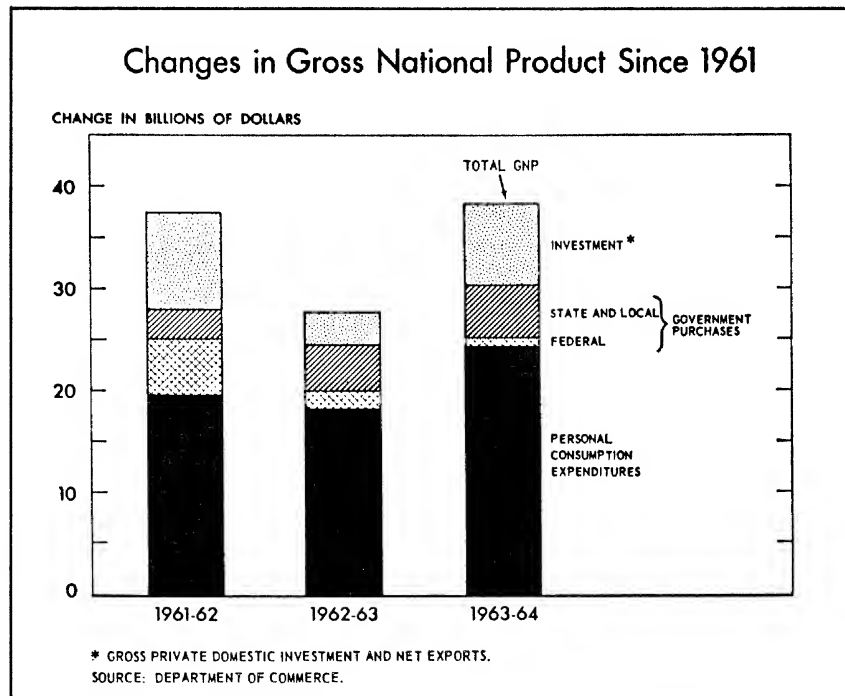
Component Part Bars

Bar charts may be used effectively to show component parts of a total as well as the total itself. The bars are subdivided into relatively few segments, each with its distinctive shading and label. The largest value is usually shaded darkest and placed at the bottom. Chart 3-9, for example, shows year-to-year changes, both in total GNP and in its major expenditure components.

Divided bars may show either absolute amounts or proportions of the total as 100 percent. If actual values are more important, the original units are shown on the scale and the bars are of varying lengths. If, on

³ An exception is the range chart, such as that in which vertical bars depict high and low stock prices. Here zero is not a factor.

Chart 3-9



SOURCE: *The Economic Report of the President*, January 1965, p. 41.

the other hand, a relative breakdown is desired, the scale is in percent, and all bars have the same length of 100 percent.

SUMMARY

Data may be presented effectively in tables and charts by following the rules suggested in this chapter. Tables offer the advantages of showing more exact values than do charts, while charts serve to attract the reader's attention and show trends and comparisons more vividly than do tables.

A table is designed to show the significant relationships of data in vertical columns and horizontal rows. The data should be classified by a definite plan—usually by qualitative, size, or time differences. The two principal types of tables are reference and summary tables. Reference tables (usually placed in the appendix) are detailed and arranged for easy reference. Summary tables are short and are arranged to emphasize important facts and comparisons in the text. A summary table may be abstracted from one or more reference tables by omitting or grouping unimportant figures, rounding off numbers to three or four significant

figures, rearranging the data for emphasis and comparability, and adding percentages, ratios, or averages.

The following principles should be followed in order to construct an effective table: (1) Confine the table to a single subject. (2) Cross classify the data so as to bring out significant relationships, but do not use more than two or three classifications. The classifications should not overlap. (3) Have the title show "what, where, when, and how classified" or let it tell the story as in a newspaper headline. (4) Include specific footnotes. (5) Make references to exact sources. (6) Arrange the table for maximum effectiveness and emphasis.

If properly executed, charts can be used effectively for both analysis and interpretation of data. In the latter case, it is particularly important that a chart have proper proportions to avoid distortion. A series of charts may be grouped consecutively with running narrative titles in order to tell a complete story graphically.

The principal forms of charts are arithmetic line charts, ratio charts, bar charts, and scatter diagrams (described in Chapter 22). The *arithmetic line chart* is the most common type, since it offers a single comparison of absolute magnitudes if the scales are correctly proportioned. Several series may be compared on a single arithmetic scale (if all are in the same unit), but it is sometimes preferable to express both as index numbers on a common base or to plot them on a ratio scale.

Ratio or *semilogarithmic* charts show relative comparisons by means of a vertical logarithmic scale, with an arithmetic time scale to picture dynamic changes. A ratio scale is constructed by plotting natural numbers at distances from the bottom line proportional to their logarithms.

Data should be plotted on one-cycle paper for maximum enlargement if the range is within the 10 to 1 ratio. The bottom of the scale should be marked 1, 2, 4, or 5 (with appropriate zeros and units) and this value multiplied by the printed scale figures to get the other values. Different scales may be used to bring series of diverse sizes and units together for easy comparison.

The ratio chart is useful for three types of comparison: (1) It shows a constant percent rate of growth as a straight line, so changes in this rate are denoted by curvature of the line, and trend forecasts can sometimes be made. (2) The relative growth or fluctuations of two curves may be compared more accurately than in arithmetic charts, since parallel lines indicate the same percent rates of change anywhere on the chart, and steeper slopes indicate higher rates. (3) Percents or ratios may be read directly from the vertical scale and applied toward further graphic analysis.

Ratio charts, however, should not be used to give a visual picture of absolute amounts or to contract a wide range of such values, nor for very simple illustrations, nor for data including zero or negative values.

Bar charts are usually preferable to two- and three-dimensional geometric forms, such as circles and solids, for showing simple comparisons. They may also be used in place of line charts for portraying a relatively few values or for representing the parts of a whole. Since bars denote size by their length, the scale should not be broken. Bars may be divided to show the changes of component parts either in absolute amounts or relative to the total as 100 percent, whichever is more significant.

PROBLEMS

The data concerning inspection of electric shavers contained in the five daily inspection reports reproduced below are to be used in preparing solutions to Problems 1-4:

SMOOTH-SHAVE COMPANY
Summary of Daily Inspection Reports

Date	Shaver No.	Machine Operator	Number of Shavers			
			Inspected	Accepted	Scrapped	Salvaged
Oct. 3	83	T. R.	2,680	2,650	30	...
	55	J. R.	1,207	1,200	7	...
	71	L. N.	2,950	2,150	800	...
	22	E. S.	1,893	1,780	113	...
	25	J. W.	1,350	1,350
4	83	T. R.	2,545	2,500	45	...
	55	J. R.	1,712	700	62	950
	71	L. N.	2,600	2,075	525	...
	22	E. S.	1,703	1,550	153	...
	25	J. W.	1,979	1,180	350	449
5	83	T. R.	1,888	1,850	38	...
	55	J. R.	1,514	1,500	14	...
	71	L. N.	2,850	2,500	350	...
	22	E. S.	1,320	1,320
	25	J. W.	383	250	28	105
6	83	T. R.	3,835	2,000	35	1,800
	55	J. R.	1,804	1,800	4	...
	71	L. N.	2,295	2,075	220	...
	22	E. S.	1,236	1,150	86	...
	25	J. W.	694	427	177	90
7	83	T. R.	2,727	2,700	27	...
	55	J. R.	1,665	1,583	82	...
	71	L. N.	2,920	2,600	320	...
	22	E. S.	1,463	1,360	103	...
	25	J. W.	1,280	1,280

1.
 - a) Prepare a table of the number of electric shavers inspected, number accepted, number scrapped, and number salvaged each day.
 - b) Prepare a table of percents from the table of part (a), above.
 - c) State some possible reasons for the day-to-day variations.
 2.
 - a) Prepare a table of the quality of work done by different operators.
 - b) As a foreman, what use would you make of this information?
 3. If some types of shavers are more complicated than others, then some should show a higher percent of scrap and salvage than others. What can you find on this question?
 4.
 - a) Prepare a table showing the percent accepted by individual operators in each of the five days.
 - b) What information does this table show that the tables prepared for Problems 1(b) and 2(a) do not give?
 5. The following statistics have been published for the United States Steel Corporation: In 1963, 18,900,000 net tons of steel products were shipped, and sales totaled \$4,129,400,000. The following year, 21,200,000 net tons were shipped for an increase of \$492.2 million in sales over the year before. In 1964 total expenses were \$3,892,600,000, a total of \$359.0 million more than the previous year. The number of employees increased from 187,721 to 199,991; they worked an average of 35.9 and 36.8 hours per week in the two years, respectively.
 - a) Present this information in tabular form, taking account of all the points of established practice in table construction. Include any desirable ratios, percents, or other derived figures.
 - b) Does your table (or tables) have unity? Explain. What degree of cross classification is present?
 6.
 - a) Present a summary table in good form condensed from a recent census publication.
 - b) Explain specifically what information the table is intended to emphasize.
 - c) List the steps taken in condensation and rearrangement.
 7.
 - a) For what purposes is graphic presentation superior to tabular presentation?
 - b) In what ways is a chart an inadequate substitute for a table?
 - c) How can the visual impression conveyed by a chart be distorted by the use of improper proportions?
 - d) Why is there danger of misinterpretation if part of the area between zero and a time series curve is omitted on an arithmetic chart?
 - e) What are the disadvantages of using different arithmetic scales in comparing several series?
 8.
 - a) Find a published chart which you consider to be correctly and effectively drawn and explain why you think it is.
-

- b) Find a published chart which you consider incorrect or ineffective and suggest changes that might improve it. Cite the exact source in each case.
- c) Find a sequence of charts which have been put together to form a connected narrative from current business or economic publications (citing the exact issues and page numbers) and list their good and bad features.
9. a) Plot the following data on three separate charts, corresponding to the three methods shown in Chart 3-3. Use 1962 as the base (100 percent) for the index numbers (i.e., divide each value by 1962 value).
- b) Explain briefly what each chart shows.

SALES AND NET PROFIT OF A SMALL
COMPANY, 1960-66

Year	Sales	Net Profit
1960	\$21,000	\$ 300
1961	28,000	500
1962	23,000	400
1963	31,000	900
1964	26,000	700
1965	47,000	1,500
1966	41,000	1,100

10. a) Plot the following data in good form on any type of chart you think suitable.
- b) Defend your choice of chart.

Year	U.S. Natural Gas Sales, Billion Cubic Feet
1940.....	2,660
1945.....	3,919
1950.....	6,282
1955.....	9,405
1960.....	12,771
1965.....	16,629*

* Estimated.
SOURCE: *Statistical Abstract*.

11. a) Discuss the relative advantages of arithmetic and logarithmic vertical scales for time series charts.
- b) How would you label the bottom and top of a printed ratio sheet for data having the following ranges: 390 to 1,400 tons; 65 to 3,200 million passenger-miles; \$0.16 to \$55.50; 89,000,000 to 180,000,000 population? How many cycles does your ratio sheet have in each case—1, 2, or 3?
12. a) Draw a ratio chart of the data given below.
- b) Interpret the facts shown by your chart.

SELECTED FARM STATISTICS, 1930-60

Year	Number of Farms (Thousands)	Farm Income (Millions of Dollars)	Number of Tractors on Farms (Thousands)
1930	6,546	11,432	920
1935	6,814	9,666	1,048
1940	6,350	11,038	1,545
1945	5,967	25,772	2,354
1950	5,648	32,482	3,394
1955	5,087	33,332	4,345
1960	3,949	37,934	4,684

SOURCE: *Historical Statistics of the United States*.

13. a) Compare the growth of two industries or companies since 1960 by plotting their annual production or sales curves on a ratio chart.
 b) Compare the percent rates of change in different years for one of the curves.
 c) Compare the relative growth of the two curves during this period.
 d) Mark a percent measuring scale on the chart. Show the percent change in each series between the first and last years by measuring the vertical difference on this scale.
14. a) Prepare a bar chart showing absolute amounts or proportions of the total, whichever is appropriate, for the following data. Arrange the automobile companies in an effective order.

 PRODUCTION OF PASSENGER CARS IN THE
 UNITED STATES
 (In Thousands of Units)

	Full Year 1964	January- August 1965
American Motors.....	394	220
Chrysler.....	1,242	901
Ford.....	2,146	1,667
General Motors.....	3,957	3,425
Total.....	7,739	6,213

SOURCE: Standard and Poor's Industry Surveys, "Autos," September 23, 1965.

- b) Which type of bar chart is better here—component parts or separate bars for each car; absolutes or relatives? Why?
 c) Justify your arrangement of companies.

SELECTED READINGS

 AMERICAN SOCIETY OF MECHANICAL ENGINEERS. *Time-Series Charts*. New York: ASME, 1960.

This manual focuses on design as it affects a chart's meaning.

FRANCIS, ELY. *Using Charts to Improve Profits*. Englewood Cliffs, New Jersey: Prentice-Hall, 1962.

Describes the use of charts as a control tool in business.

HUFF, DARRELL. *How to Lie with Statistics*. New York: W. W. Norton, 1954. Chapters 5, 6, and 9 illustrate some common misuses of charts.

SCHMID, CALVIN F. *Handbook of Graphic Presentation*. New York: Ronald Press, 1954.

A complete, readable treatment of graphic techniques and their use in designing the principal types of charts, with many illustrations.

SPEAR, MARY E. *Charting Statistics*. New York: McGraw-Hill, 1952.

A book on practical graphic presentation, depicting many types of charts and their uses in economics.

U. S. DEPARTMENT OF AGRICULTURE. Agriculture Handbook No. 128. *Graphic Analysis in Agricultural Economics*. Washington, D.C.: Superintendent of Documents, 1957.

Applies graphic methods of analysis to frequency distributions, time series, correlation, linear programming, and many other fields of statistics.

WALLIS, W. ALLEN, AND ROBERTS, HARRY V. *Statistics, A New Approach*. New York: The Free Press, 1956.

Chapters 6 and 9 discuss the art of organizing data and the use of tables to reveal the association of different series.

ZEISEL, HANS. *Say It with Figures*. 4th ed. New York: Harper & Row, 1957.

An advanced book covering problems of classification, methods of numerical presentation, and principles of making tabulation decisions.

4. ANALYSIS OF DATA: RATIOS AND FREQUENCY DISTRIBUTIONS

STATISTICAL METHODS deal with the collection, presentation, analysis, and interpretation of data. Chapters 2 and 3 have described the methods of collecting statistical information and presenting the results in tables and charts. Beginning in this chapter, we take up the principal methods of analyzing and interpreting data.

The first step is to reduce large masses of raw figures to a simple form. As noted in Chapter 3, such data may be classified in three ways: by qualitative characteristics, by size, and by time. (These classifications were illustrated by unemployment rates in Table 3-1.) In Chapter 4 we first discuss ratios as a simple method of comparing qualitative data, and then frequency distributions as a means of summarizing data classified by size. Chapters 18 to 21 will be devoted to time series analysis.

The criteria used in classifying qualitative data are often called *attributes*. An attribute is a characteristic that can be divided into two or more categories, such as the "yes" or "no" responses on a questionnaire; "defective" or "good" in describing the quality of a product; or a classification of employees as executives, office workers, and factory workers. However, attributes usually refer to only two categories (e.g., factory workers and other employees), and ratios are used to compare just two categories, such as the ratio of factory workers to total employees.

Data classified by size or time, on the other hand, are called *variables*. Thus, a size classification might be the number of unemployed classified by age of workers, where age is the variable. Variables classified by size may be grouped into frequency distributions, and averages and measures of dispersion may be computed to summarize their characteristics, as described in the latter part of this chapter and in Chapters 5 and 6.

RATIOS

A ratio or proportion is an extremely useful and simple device for comparing two attributes or qualitative characteristics. Thus, it is usually more significant to report the unemployment *rate* (i.e., the ratio of unemployed to total labor force) than simply to give the total number of unemployed. Ratios are also useful in comparing groups of variables classified by size, such as citing the percentage of factory workers who earn less than \$2.50 an hour, even though the basic data are classified by size of hourly earnings. This section describes how to construct ratios that are accurate and meaningful for economic analysis and how to interpret them.

The ratio of one number to another is a fraction in which the first number is the numerator and the second number is the denominator or *base*. Often, the two numbers are expressed in the same units (e.g., dollars) as in a company's ratio of net profits to sales.

Various terms are used for ratios in which the terms are measured in different kinds of units. Thus, the birth *rate* is the number of births per thousand population; *density* of population is the number of persons in a region divided by its area; *per capita* national debt is the ratio of total debt to the number of persons in the country.

It is important to present a statistical ratio in such a way that the reader understands exactly what quantities are being compared, particularly when the units of the two terms of a ratio are different.

Selecting the Numerator and Base

The quantities selected for a statistical ratio should be related to each other in such a way that their ratio will be most meaningful for the problem at hand. Often, one or both of the quantities can be adjusted or refined so as to exclude any extraneous factors that would obscure the direct relationship between them. For example, the ratio "farm income per acre" in a given state would be more meaningful if the denominator were adjusted to exclude forests, deserts, and other nonfarm land, to provide the ratio "farm income per acre of *arable* land."

In the same way, safety departments of manufacturing plants get an accident rate for each department by taking the ratio of employees injured to total number of operating employees, excluding office workers. Both the numerator and denominator are adjusted further in order to facilitate the study of accidents. The resulting ratio, known as the accident severity rate, is the number of days' work lost through acci-

dents¹ divided by the number of equivalent full-time days worked per week or month.

The study of deaths in automobile accidents furnishes another example of the need for refining the figures used in computing ratios. Table 4-1, row 1, shows that the number of persons killed in motor vehicle accidents increased 37 percent between 1950 and 1964. These figures suggest that the "automobile menace" is increasing. The increase may be due to the growth of population, however, so the number of deaths per 100,000 population has been computed, as shown in row 2. This ratio has increased by only 8 percent. However, accidents are related

Table 4-1
FATALITIES IN MOTOR VEHICLE ACCIDENTS, 1950 AND 1964

	1950	1964	Percent Change
1. Persons killed in motor accidents.....	34,763	47,700	+37
2. Deaths per 100,000 population.....	23.0	24.9	+8
3. Deaths per 10,000 motor vehicles.....	7.1	5.5	-23
4. Deaths per 100,000,000 vehicle-miles..	7.6	5.7	-25

SOURCE: National Safety Council, *Accident Facts*, 1965, p. 59.

more directly to the number of motor vehicles, which have increased more rapidly than the total population. The number of deaths per 10,000 motor vehicles, therefore, is shown in row 3. Now we see a 23 percent *decrease* in this refined ratio. Finally, traffic deaths are related still more specifically to the number of vehicle-miles driven, and the average car was driven more miles in 1964 than in 1950. The number of deaths per 100,000,000 vehicle-miles is shown in row 4. The decrease is now 25 percent. The more refined ratio therefore shows a substantial gain in safety, when the increased number of cars and mileage driven are taken into account, whereas the actual fatalities and the crude per capita ratio (rows 1 and 2) indicate just the opposite conclusion.

¹The number of days' work lost can be counted for temporary accidents but not for death or permanent disability. Consequently, standards have been established for each type of accident. Thus, according to one standard, 6,000 days are allowed for death, 4,000 days for loss of an arm, 1,200 days for loss of a thumb and one finger, etc., U.S. Bureau of Labor Statistics Bulletin No. 234, *The Safety Movement in the Iron and Steel Industry*, p. 278.

Which Item to Choose as Base

The base or denominator of a statistical ratio is always a standard with which the numerator is being compared. The numerator is the quantity on which the inquiry is focused; the denominator provides the basis for comparison. The following rules may be useful in selecting the base:

1. In comparing a part and the whole, the whole is always the base. Example: net profits to sales ratio = $\text{net profits} \div \text{sales}$.
2. In time comparisons of like items, the prior event is almost always taken as the base. Example: this year's sales as a percent of last year's.
3. In comparing a cause and effect or an independent event with one at least partly dependent on it, the cause or the independent item is nearly always the base. Example: price-earnings ratio of a common stock = $\text{price} \div \text{earnings}$. (Exception: stock yield = $\text{dividend} \div \text{price}$.)

When either of two items is equally logical as a base, custom often determines the choice. Example: rate of inventory turnover = $\text{sales} \div \text{inventory}$.

The Number of Units in the Base. The base may be expressed as a single unit, 100 units, or some other multiple of ten, depending on which is customary or most effective. Thus, the national debt of \$1,627 per capita is expressed in terms of *one* denominator unit, or one person; an interest rate of 4 percent means \$4 for every \$100 deposited, whereas the death rate may be reported as 9.0 per *thousand*. As shown in Table 4-1, the National Safety Council reports motor vehicle deaths per 10,000 motor vehicles, per 100,000 population, and per 100,000,000 vehicle-miles. The larger numbers are used as a base so that the numerator can be reported mainly as a whole number rather than as a decimal fraction.

However, most ratios used in statistics are expressed in terms of *percents*, provided they compare identical units; comparisons of unlike quantities are expressed in terms of the base unit, such as motor vehicle deaths per 10,000 motor vehicles.

Cautions in the Use of Ratios

Many of the errors in the use of ratios spring from failure to express the meaning of ratios correctly. Thus, an advertisement reads: "In January 1955, there were only 330 [—Rent-a-Car] Offices. Today we opened our 1000th station—a growth of over 300%. . . ." The increase from 330 to 1,000 was 670, a growth of only 203 percent.

A further error in the use of percents should be noted. The difference between two percents, often called "percentage points," must not be interpreted as a percent change. Thus, it is incorrectly stated that "average weekly earnings of factory workers in 1964 were 25 percent above the 1957-59 level, but in 1965 they had risen to 37 percent, a 12 percent increase." These are both percents of the same base period, the 1957-59 level, but the percent change is obtained by dividing the increase of 12 percentage points by the base level of 125, an increase of $9\frac{1}{2}$ percent, not 12 percent.

As has already been indicated, the base item in time ratios is practically always the earlier period. Failure to observe this rule leads to still further confusion in the expression of percentage increase or decrease, as illustrated in the following newspaper headline: "Liquor Prices Cut 200 Percent in Price War." Whatever the former price may have been, however, a cut of 100 percent would reduce it to zero. Hence, any greater decline would mean that the retailers were paying the purchasers to take their wares! What probably happened was that liquor formerly selling at \$6.00 per quart was cut \$4.00 and placed on sale at \$2.00. Dividing \$4.00 by \$2.00, the later price, gives 200 percent; but this is the percentage by which the past exceeded the present, not the percentage decrease. The correct practice would have been to use the original price as the base of the ratio. That is, the cut was $\$4.00 \div \$6.00 = 66\frac{2}{3}$ percent.

Finally, ratios should not be used if the original number used as base is very small. The report that 25 percent of the bank tellers in a town have been indicted for embezzlement would be misleading if there were only four tellers to begin with. Similarly, a 1,000 percent increase in profits over last year would hardly be significant if last year's profits totaled \$1.

Whenever possible, the data from which ratios have been derived should be shown with the ratios. The reader is rightly skeptical in accepting any statement of relationships that he cannot verify by making the computation himself. Sometimes additional relationships can be derived from a given set of data. If the original data are not shown, the reader is prevented from working out ratios which may be of more interest to him than those selected by the author.

FREQUENCY DISTRIBUTIONS

Many types of data are classified according to size. Examples are rents paid for houses, population by age groups, and wages of workers. In each case the original data are values of a *variable* (e.g., rent, which

varies from house to house) which will be called X . These are classified by assigning each value to the size class or *class interval* to which it belongs. The number of values of X in each interval is the *frequency*, and the whole table of frequencies is a *frequency distribution*.

A *frequency distribution* therefore is a table in which values of a variable are classified according to size. It is a valuable device for summarizing unwieldy figures, so that a maximum of information can be presented with a minimum of detail.

Variables may represent either discrete or continuous data. Discrete data have distinct values, with no intermediate values. Thus, the number of children in a family can be two or three, but not 2.7. Continuous data can have any values over a range, such as the exact heights of men. However, continuous data are often treated as being discrete, such as when heights are rounded to the nearest inch, and a man's height is reported at either 5 ft. 10 in. or 5 ft. 11 in. but not at any intervening value.

In order that the analysis of data may be meaningful, it is necessary that they be *homogeneous*, that is, sufficiently alike to be comparable for the purposes of the study.

Homogeneity may be illustrated by a study of gasoline prices in Rockford, Illinois, conducted for the Standard Oil Company of Indiana. Here, the prices for the regular grade at major-brand service stations varied from 30.3 to 31.7 cents a gallon, while the prices at private-brand or "cut-rate" stations varied from 27.4 to 29.9 cents. Hence, each of these homogeneous groups was analyzed separately. If all stations had been combined, the resulting distribution would have been *heterogeneous* and would have concealed important differences in pricing policy of the two types of station.

The Array

Sometimes it is convenient to arrange the values of the variable in an *array*, as a preliminary step. An array is a listing of values arranged in order of *size*—either from smallest to largest or vice versa. The values can either be listed individually or summarized on a tally sheet.

Table 4-2, for example, shows the overall dimension of 63 gears, taken from a quality control measurement. The raw data in panel A are too awkward to handle directly, so they have been combined in an *array* in panel B by means of a tally sheet.

This array not only shows the data in simpler form than in panel A but reveals at a glance certain salient characteristics—the highest and lowest values and the most frequent size (.4250 in.). Also, in this simple case where no further grouping of values is needed, the array is

Table 4-2

RAW DATA AND ARRAY
DIMENSIONS OF 63 GEARS AS ILLUSTRATED, INCHES

A

. 4260	. 4260	. 4250
. 4240	. 4240	. 4240
. 4255	. 4245	. 4250
. 4250	. 4245	. 4265
. 4235	. 4260	. 4255
. 4265	. 4245	. 4255
. 4255	. 4260	. 4250
. 4260	. 4245	. 4240
. 4250	. 4240	. 4245
. 4245	. 4240	. 4255
. 4235	. 4250	. 4260
. 4240	. 4260	. 4245
. 4250	. 4250	. 4260
. 4245	. 4250	. 4255
. 4255	. 4255	. 4250
. 4250	. 4245	. 4245
. 4255	. 4250	. 4255
. 4265	. 4260	. 4260
. 4270	. 4250	. 4255
. 4255	. 4235	. 4250
. 4250	. 4265	. 4245

Technical drawing of a mechanical component, likely a gear or pulley, showing a cross-section with teeth and a central hub. Dimensions are indicated: 422 and 421.

B

MARCHANT CALCULATORS, INC.

SQ Date 9/6 Insp. BA Qty 63

Part No. 17110 Dim. 422/428

. 4290	1111
85	
. 4280	
75	
. 4270	1
65	
. 4260	THL THL
55	THL THL 11
. 4250	THL THL THL
45	THL THL 1
. 4240	THL 11
35	111
. 4230	
25	
. 4220	
15	
. 4210	
05	
. 4200	

SOURCE: Marchant Calculators, Inc., *Statistical Quality Control*.

already in the form of a usable frequency distribution, with class intervals .0005 in. wide—the number of marks opposite each dimension indicating the frequency with which this measurement occurred.

Grouping Data into Classes

Most types of data, however, have so many different values that an array is excessively detailed. The figures must then be grouped into a manageable number of classes. The methods for doing this are illustrated below with data adapted from a survey of straight-time hourly earnings of 214 apprentice machine tool operators in machinery manufacturing plants in an eastern city. Studies of this type are needed for industrial relations analysis, labor-union wage negotiations, and many aspects of welfare economics.

Table 4-3 presents an array of these hourly earnings in the form of a tally sheet, with the number of operators at each earnings level noted in the column headed "*f*" (for frequency). This table still has too many separate values for easy analysis and presentation, so the data are grouped as shown in Table 4-4. For this purpose, class intervals 10

Table 4-3

MORE DETAILED ARRAY
STRAIGHT-TIME HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL OPERATORS
IN MACHINERY MANUFACTURING PLANTS IN AN EASTERN CITY
(In Dollars per Hour)

EARN- INGS	OPERATORS		EARN- INGS	OPERATORS		EARN- INGS	OPERATORS	
	Tally	f		Tally	f		Tally	f
2.30		1	2.55		5	2.80		5
2.31			2.56		6	2.81		1
2.32		1	2.57		3	2.82		
2.33			2.58		4	2.83		
2.34			2.59		5	2.84		
2.35		2	2.60		11	2.85		1
2.36		2	2.61		4	2.86		1
2.37			2.62		3	2.87		1
2.38		3	2.63		20	2.88		
2.39		2	2.64		2	2.89		
2.40		7	2.65		9	2.90		
2.41		1	2.66		2	2.91		
2.42			2.67		3	2.92		
2.43		1	2.68		2	2.93		
2.44		5	2.69		3	2.94		
2.45		4	2.70		13	2.95		
2.46		3	2.71		3	2.96		
2.47		5	2.72		6	2.97		1
2.48		3	2.73		1	2.98		1
2.49		2	2.74		3	2.99		
2.50		12	2.75		11	3.00		
2.51		5	2.76		5	3.01		
2.52		1	2.77		1	3.02		1
2.53		12	2.78			3.03		
2.54		2	2.79		2	3.04		1

Table 4-4

FREQUENCY DISTRIBUTION
HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL OPERATORS

Hourly Earnings	Midpoint	Number of Operators f	Percent of Operators
\$2.25 and under \$2.35	\$2.30	2	1
\$2.35 and under \$2.45	2.40	23	11
\$2.45 and under \$2.55	2.50	49	23
\$2.55 and under \$2.65	2.60	63	29
\$2.65 and under \$2.75	2.70	45	21
\$2.75 and under \$2.85	2.80	25	12
\$2.85 and under \$2.95	2.90	3	1
\$2.95 and under \$3.05	3.00	4	2
Total		214	100

cents wide were chosen, beginning with \$2.25 as the lower limit of the first interval. The *class interval* is the range of values for each class. This is the difference between the lower limits, or upper limits, of two consecutive classes.

The reasons for this choice of intervals are as follows: The number of classes (eight) is large enough to show the general distribution of earnings and small enough to simplify analysis and presentation. The class limits (\$2.25, \$2.35, etc.) are multiples of 5 cents, which are simple round numbers, while the midpoints (\$2.30, \$2.40, etc.) are at the popular rates at multiples of 10 cents. This permits easy interpretation and minimized errors of grouping. Finally, the intervals (\$2.25 and under \$2.35, etc.) are defined clearly and unambiguously. These principles are discussed below.

Number and Width of Class Intervals

In general, it is advisable to divide the data into from 6 to 15 classes. If the number of classes is too small, important characteristics of the data may be concealed by grouping in intervals that are too broad. At the other extreme, it is rarely necessary to preserve so much detail that more than 15 classes are needed.² Also, if there are too many classes, there may be a confusing zigzag of frequencies, and some classes may contain no values of *X* at all—particularly if the total number of items is small. This is the case in Table 4-3, which lists 75 one-cent intervals.

Once the approximate number of classes has been chosen, the exact number is determined by the width of the interval. This interval is usually selected as a convenient round number located so that clusters of data occur at its midpoints, as described in the next section. Thus, in Table 4-4, earnings tend to cluster at multiples of 10 cents, so we have used \$2.30, \$2.40, etc. as class midpoints, and the 10-cent interval gives us eight classes. There are also minor clusters at odd multiples of 5 cents, however, so we could have placed all of these points of concentration at midpoints by using intervals 5 cents wide beginning with "2.275 and under 2.325." It is doubtful, however, whether the slight increase in accuracy is worth the use of odd figures as class limits and the additional work required by the larger number of classes.

Choice of Class Limits and Midpoints

The midpoint of a class interval is halfway between its limits. The exact location of the class limits depends on the method of reporting the

² Some writers, however, suggest from 6 to 15 classes for presentation but from 15 to 25 classes for accuracy in computations.

original data and subsequent rounding, if any. For example, in population censuses, ages are reported to the *last* birthday. Here, the five-year interval "20–24" includes all persons from their twentieth birthday to the eve of their twenty-fifth birthday. In this case, therefore, the midpoint is halfway between 20 and 25, or 22.5. On the other hand, when ages are rounded off to the *nearest* birthday, as in life insurance practice, the class interval "20–24" is interpreted as 19.5 up to, but not including, 24.5. Thus the midpoint is 22.

The midpoint of an interval in a frequency distribution is used to represent the average value of all the items in the class. This usage

Table 4-5
METHODS OF DESIGNATING CLASSES
FOR BEGINNING SALARIES OF COLLEGE GRADUATES
(In Dollars per Month)

A Possible Value Limits	B Upper Limit Excluded	C Overlapping	D Midpoint
425–474	425 and under 475	425–475	450
475–524	475 and under 525	475–525	500
525–574	525 and under 575	525–625	550
etc.	etc.	etc.	etc.

involves *errors of grouping*, which are similar to *errors of rounding off* numbers in general. For example, in rounding off the age 22.4 to 22, the error is 0.4. It is important to minimize the errors of grouping by locating the midpoints of the intervals at any *points of concentration* around which values tend to cluster. Otherwise, any averages or other measures computed from the frequency distribution would be biased. Thus, if monthly salaries paid college graduates were set by a company at multiples of \$50—say \$500, \$550, \$600, etc.—and they were reported in a frequency distribution with classes "\$500 and under \$550," etc., so that the midpoint of \$525 was used to represent salaries that were all actually \$500, a computed average would overstate the true value by \$25. If midpoints are located at points of concentration, however, errors of grouping are not serious, because the errors in different classes tend to offset each other.

Designation of Classes. The class limits should be stated precisely to avoid ambiguity. For example, suppose we wish to classify the beginning monthly salaries of college graduates in intervals of \$50, with midpoints at multiples of \$50. We could list the classes in any of four ways, as in Table 4-5.

The listing in column A is appropriate for discrete data, with salaries reported to the nearest dollar. (However, if salaries were reported in dollars and cents, the limits would have to read "425.00–474.99" etc. to include all values.) The listing in column B is suitable for either discrete or continuous data and is usually the clearest method of designating classes. On the other hand, one should avoid the listings shown in columns C and D since they are ambiguous; it is not clear in what class the limiting values such as \$475 and \$525 fall.

Uniformity in Width of Class Intervals

It is highly desirable that all intervals used in a frequency distribution have the same width, because frequencies are easier to interpret and averages are easier to compute. Intervals of varying width are confusing and awkward to use in analysis. Unequal intervals are often necessary, however, in order to cover a wide range of data, as in the following grouping of annual incomes:

Under \$2,000	\$ 6,000–\$ 9,999
\$2,000–\$3,999	\$10,000–\$19,999
\$4,000–\$5,999	\$20,000 and over

In such cases, it is also rather common to have *open-end classes* at the extremes, with the lower limit of the smallest class and the upper limit of the largest class not shown. For example, "under \$2,000" and "\$20,000 and over." This open-end type of frequency distribution is sometimes needed to include a few extremely large or small values without adding a number of extra classes. The sum of the values in such open-end classes should be indicated, if possible, to aid in computing averages and other summary measures.

Relative Frequency Distributions

It is often desirable to show each frequency as a relative or percentage of the total, as shown in the last column of Table 4–4.

The use of percentages has four advantages: (1) It permits comparisons of the individual frequencies with each other and with the total on a common 100 percent base. (2) It facilitates comparisons between two frequency distributions having different numbers of items, provided they have identical class limits, as in Chart 4–3. (3) It permits one to make inferences from sample data regarding the population, provided the sample is carefully selected. For example, it might be inferred from Table 4–4 that about 29 percent of *all* Class A machine tool operators in the area earn from \$2.55 to \$2.65 an hour. (4) It provides a basis

for estimating probabilities. Thus, if we take an operator at random, we can say that the probability is .29 that he will earn from \$2.55 to \$2.65 an hour. The use of relative frequencies to estimate probabilities is described in Chapter 7.

CHARTS OF FREQUENCY DISTRIBUTIONS

A frequency distribution may be presented as a chart designed to picture its main characteristics. To construct such a chart, measure the variable X along the horizontal scale and label either the class limits or midpoints. Then, at the midpoint, plot the frequency of the class on the vertical scale (assuming classes of equal width). Both the horizontal and vertical scales are the ordinary arithmetic type. The vertical scale must always begin at zero, but the horizontal scale need only include the range of X values and one extra interval at each end. The two most common frequency diagrams of sample data are the histogram—a vertical bar chart—and the frequency polygon—a line chart. The smooth frequency curve, used to describe the distribution of values in a population, is discussed later in this chapter.

The Histogram

A histogram is a set of vertical bars whose *areas* are proportional to the frequencies represented. When the class intervals, or bar widths, are equal, the *height* alone can be used to represent the frequency in that class. The height of the bar thus shows frequency *per unit width*. The bars may be separated to show the breaks in discrete data, but they should adjoin to represent continuous data.

In Chart 4-1, for example, the histogram represents the earnings of the 214 machine tool operators listed in Table 4-4. This chart shows at a glance how the earnings are distributed.

The class which contains the greatest concentration of earnings figures is called the *modal class*. It stands out in the chart as the tallest bar. On either side, the bars taper off in height, showing that the farther the earnings are from the modal class, the fewer are the number of workers. Many types of economic data have this type of distribution—approximately symmetrical with a modal class near the center.

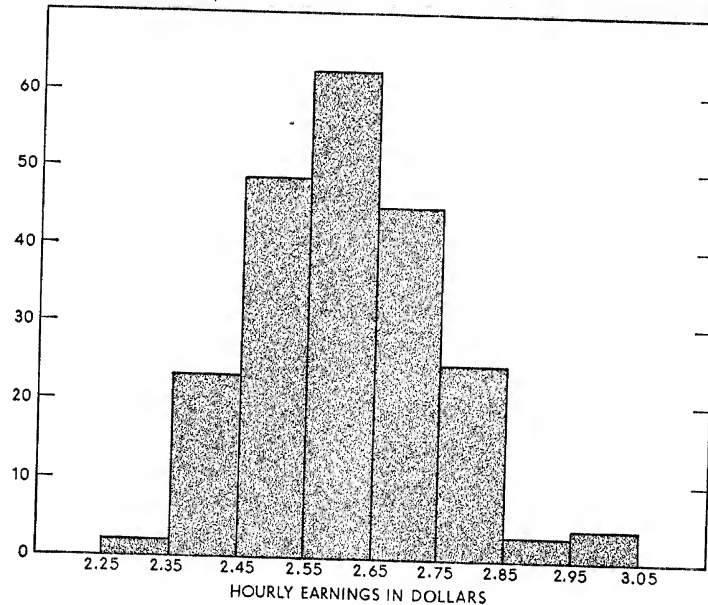
If there are two separate modal classes in a histogram, the data may prove to be heterogeneous (e.g., foremen might have been included with operators). In this case, the figures should be separated into homogeneous groups before being analyzed.

The height of each bar of a histogram is equal to the frequency of the

class when intervals are equal in width; but when the width varies, frequency is represented only by *area* rather than by height. Thus, in Chart 4-1, if the seven operators in the two classes \$2.85 to \$3.05 were combined into a single class, the height of this bar should be plotted as $7 \div 2 = 3\frac{1}{2}$, so that it would have the same area as the two right-hand bars shown. If the combined bar were drawn with a height of 7, it would double the apparent number of these highly paid workers.

Chart 4-1

HISTOGRAM
HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL OPERATORS
NUMBER OF OPERATORS (f)

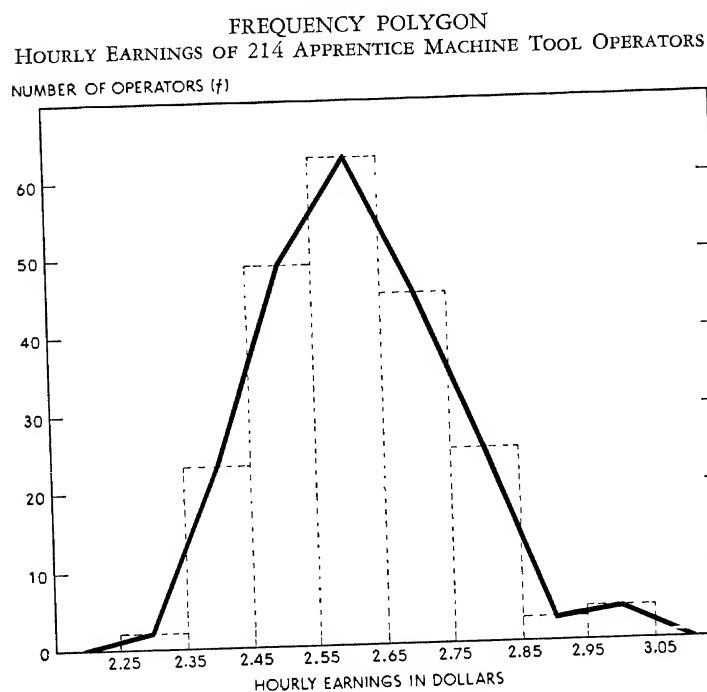


The Frequency Polygon

The *frequency polygon* is a line chart plotted on the same scales as a histogram. To draw a polygon, plot each frequency on the vertical scale over the midpoint of the interval on the X axis (assuming classes of equal width). Then connect these points with straight lines, and extend them to an interval of zero frequency at each end.

Chart 4-2 shows the frequency polygon in comparison with the equivalent histogram (which is lightly blocked in as background). The frequency polygon (including the base line) encloses an area equal to

Chart 4-2



that of the histogram,³ although the areas in individual classes are shifted slightly from the classes to which the frequencies belong.

Histograms versus Frequency Polygons

The histogram has the following advantages over the frequency polygon: (1) the area within each bar represents the exact number of values in a class; (2) the individual classes stand out more clearly than in a frequency polygon; and (3) separated bars may be used to emphasize gaps in a discrete distribution.

Frequency polygons have these advantages: (1) they are simpler than bar charts, having fewer lines; (2) they resemble the smooth curve which describes a population of continuous data better than does the histogram; and (3) they are simpler for comparing two frequency diagrams.

Histograms are usually preferable when classes are few; frequency polygons when classes are numerous. Either type of chart, however, can ordinarily be used.

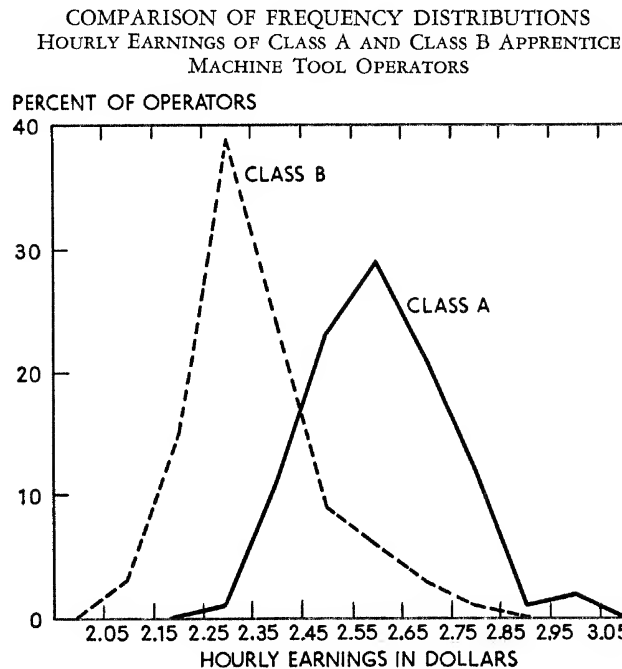
³ This follows from the fact that each pair of adjoining triangles formed by the top lines of the polygon and the histogram in Chart 4-2 are equal in area. Similar areas are not equal, however, when intervals are of unequal width.

Frequency charts have an advantage that is characteristic of all charts—they provide a quick and simple method of summarizing and presenting facts. An apparel manufacturer, for instance, can use this type of diagram in controlling his purchases and inventory. From his sales records he can prepare frequency charts showing the sizes of clothing, shoes, and other merchandise characteristic of his customers, to serve as guides in purchasing and inventory control.

Comparison of Two Frequency Distributions

Two frequency distributions can best be compared by plotting their relative frequencies as polygons on the same scales. To illustrate, Chart 4-3 compares the earnings of our Class A apprentice machine tool operators with those of Class B apprentices. The frequencies are expressed as percentages of their respective totals. Comparison of the two curves shows that (1) Class A operators earn more than Class B operators for the most part; (2) the most common earnings rates are in the \$2.25 to \$2.35 bracket for the Class B workers, as compared with \$2.55 to \$2.65 for the Class A men; and (3) there is a much greater concentration of Class B earnings than Class A earnings in these modal classes, as shown by the relative heights of the two curves.

Chart 4-3



CUMULATIVE FREQUENCY DISTRIBUTIONS

Sometimes one needs to know the answers to questions such as "How many operators earn less than \$2.75 an hour?" If so, it is convenient to add the frequencies cumulatively, beginning at either end, and list the resulting subtotals in a *cumulative frequency distribution*, as in Table 4-6, columns 3 and 4.

Table 4-6

CUMULATIVE FREQUENCY DISTRIBUTIONS
HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL OPERATORS

(1) Hourly Earnings	(2) Number in Class with Lower Limit Shown	(3) Number Earning Less	(4) Number Earning as Much or More
\$2.25	2	0	214
2.35	23	2	212
2.45	49	25	189
2.55	63	74	140
2.65	45	137	77
2.75	25	182	32
2.85	3	207	7
2.95	4	210	4
3.05	0	214	0
Total	214		

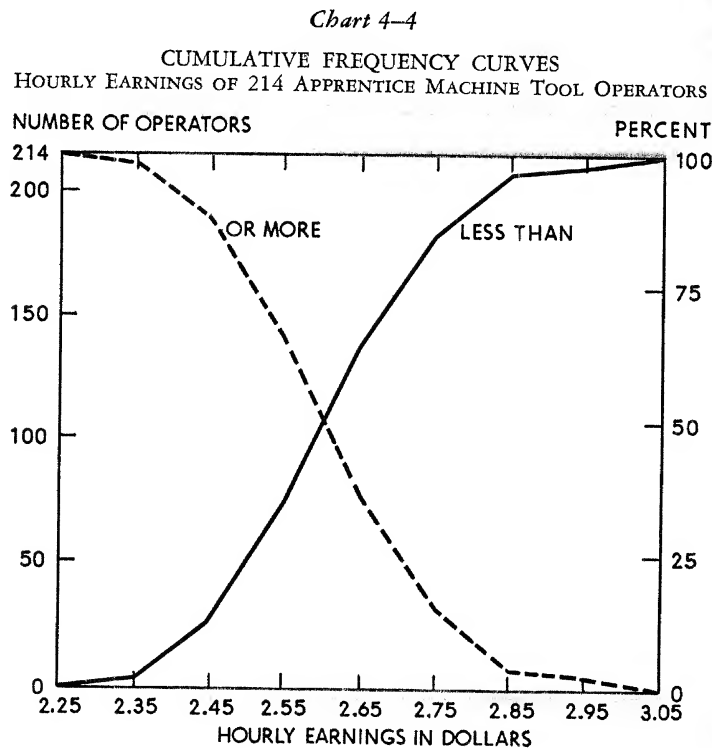
SOURCE: Table 4-4.

The table shows at a glance how many operators earn *less than* any amount listed, or that amount *or more*. Thus, 182 operators earn less than \$2.75, while 32 earn \$2.75 or more. Columns 3 and 4 could also be expressed as percents of the total number of operators (214) for better comparability with other groups or for making inferences about a larger population.

The graph of a cumulative frequency distribution is called a cumulative frequency curve or an *ogive* (pronounced ō'jīve), because its shape resembles that of an ogive or rib of a Gothic arch. The data in Table 4-6 are graphed in Chart 4-4. The percent scale at the right is made so that 100 percent corresponds to 214 operators on the left-hand scale. The ogives then show graphically what number or percent of the operators earn less than the amounts listed in Table 4-6, and what percent earn those amounts or more.

In addition, the ogives permit easy interpolation for finding values

between the plotted points. For example, the upward ogive shows that 25 percent, or about 53 operators, earn less than \$2.51, while the downward ogive shows that 25 percent earn \$2.70 or more. The intersection of the two curves at the 50 percent horizontal line indicates that about half the workers earn \$2.60 or less, and half more. These three



earnings figures are the quartiles and median, discussed in the next chapter.

The same percents can be used to make inferences about *all* comparable machine tool operators, provided the group of 214 is a good sample of the population. In this case, the sample was carefully selected so it can be inferred that about 25 percent of *all* such operators earn less than \$2.51 etc.

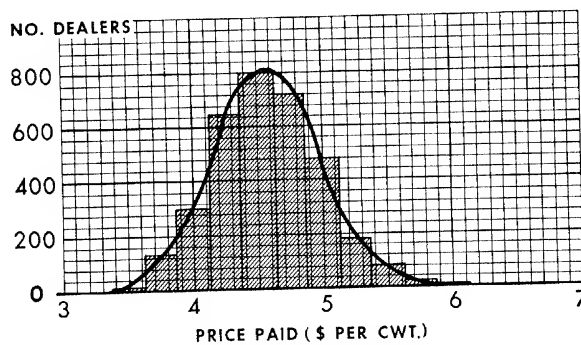
An ogive can also be drawn as a smooth *curve* through the plotted points, with the aid of a French curve, rather than as a series of straight lines. The use of the curve implies *gradual change* in degree of concentration—often a more realistic assumption than that the values are *uniformly* distributed over each class interval.

FREQUENCY CURVES

A smooth curve can be drawn to portray the frequency distribution of a *population* of continuous data. This is the limiting form of either the histogram or frequency polygon as the number of values in the sample becomes infinitely large and the class intervals become infinitely small. A frequency curve smooths out sampling errors which are particularly evident in small samples—and provides a frequency value for *every* value of X , rather than just one value for each class interval. Smooth curves cannot be used, however, for data that cluster around certain values, such as the machine tool operators' earnings in Table 4-3.

Chart 4-5

FREQUENCY CURVE FITTED TO SAMPLE DATA
LAYING MASH: PRICES REPORTED BY FEED DEALERS, SEPTEMBER 1949



SOURCE: Frederick V. Waugh, *Graphic Analysis in Economics*, U.S. Department of Agriculture, Agricultural Handbook 128 (1957), p. 3.

Chart 4-5 shows a histogram of the prices charged by 3,395 dealers throughout the United States for laying mash. The height of each bar shows the number of dealers reporting prices within that price interval. A smooth curve has been drawn by Frederick V. Waugh of the U.S. Department of Agriculture to show "the general nature of the distribution." Such curves may be fitted either graphically, on a judgment basis, or by mathematical methods. A careful study of the data is necessary in either case to assure a realistic fit. In the graphic method, the curve should be drawn in such a way that *the area cut from each bar is approximately equal to the area added to that bar by the curve*. Chart 4-5 deviates from this rule slightly in the case of the two tallest bars in order to follow a "normal curve." This type of curve is described below.

Types of Frequency Curves

Some common types of frequency curves are illustrated in Chart 4-6. The most important is the bell-shaped *normal curve* shown in Charts 4-5 and 4-6, panel A. This curve describes the distribution of many kinds of measurement in the physical, biological, and social sciences. Thus, the prices of laying mash in Chart 4-5 vary with freight rates, differences in ingredients, dealers' markup, etc., but nevertheless form a nearly normal distribution. The normal curve is particularly important, moreover, because it reflects variations due to *chance*, such as the errors in random sampling. This curve will be used in the later chapters in studying the reliability of sample measures and in making inferences about populations.

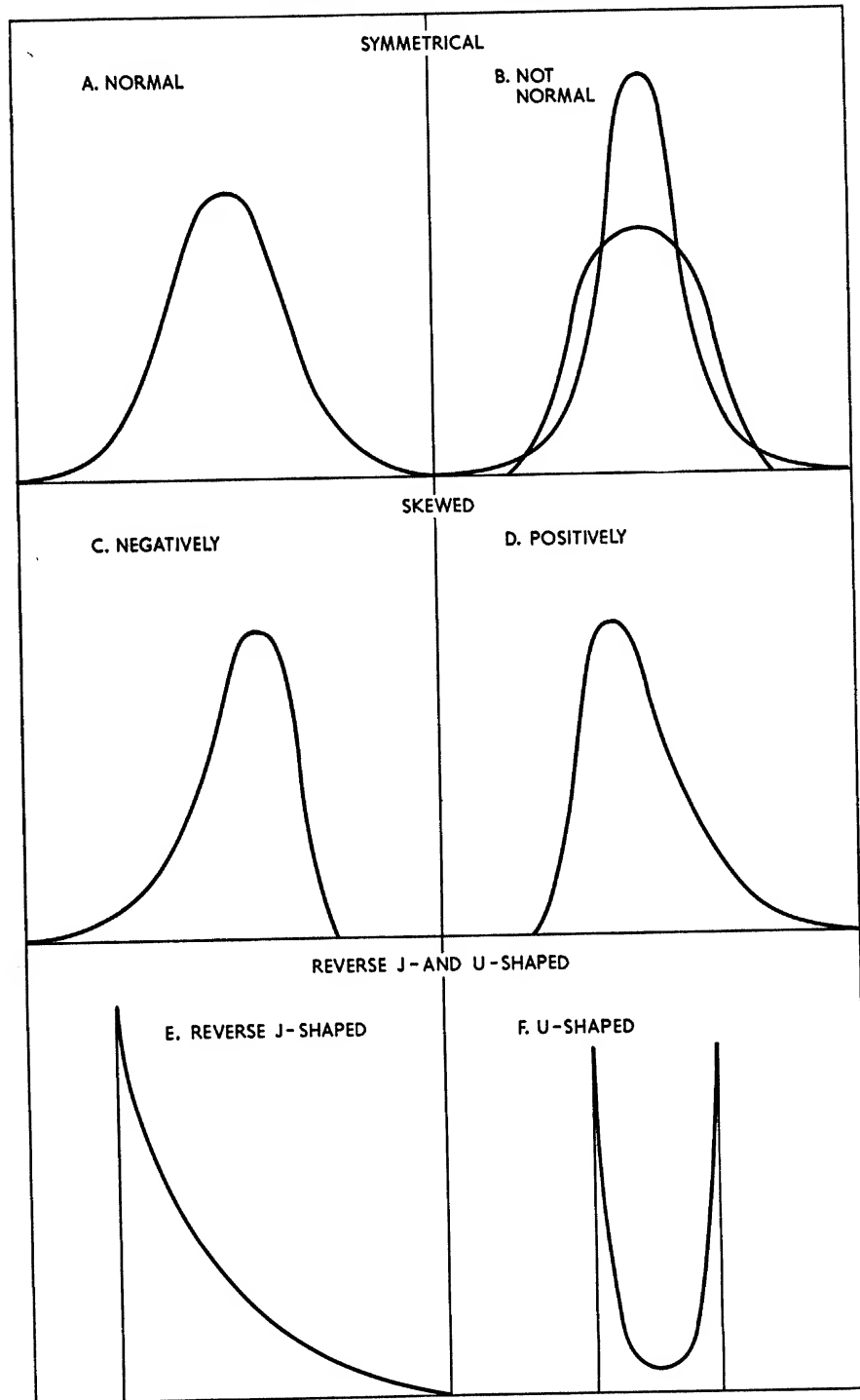
The two curves in panel B of Chart 4-6 are symmetrical like the normal curve, but one is more peaked, with longer tails; the other is more squat, and with shorter tails than the normal curve. The peaked curve might represent prices of gasoline in a city where most service stations charged about the same price, but a few prices were widely scattered. The squat curve would show that prices were distributed more evenly over a limited range, but without being concentrated at one point.

Curves C and D represent distributions that also have a "central tendency," as shown by the peak near the center of the curve, but the two branches of the curve are unequal or "skewed." Curve C, with the longer branch to the left in the negative direction, is called "skewed to the left" or "negatively skewed." This type of curve commonly results from a distribution having a fixed upper limit but a more remote lower limit, as in the case when test scores cluster closer to the perfect score than to zero. Curve D, which is skewed to the right, or positively skewed, is the most common type encountered in business and economic data. Distributions of personal earnings, commodity prices, or assets of companies, for example, tend to cluster closer to the lower limit of zero than to the indefinite upper limit. An appropriate test given to a uniform group of job applicants might produce a symmetrical grade distribution, whereas a more difficult test would produce scores lower on the average and skewed to the right, while an easier test would produce scores higher on the average and skewed to the left.

Curves E and F are less common. The reverse J-shaped curve occurs in some distributions, such as income tax payments, where the smallest returns are most numerous and the number of returns (on the Y axis)

Chart 4-6

TYPES OF FREQUENCY CURVES



drops off sharply at first and then more gradually as the size of payment (on the X axis) increases. The U curve may be illustrated by the number of houses classified by percent of mortgage debt to house value, where many houses have no debt or a heavy debt, while relatively few have a middle-sized debt in relation to house value. The averages and measures of dispersion discussed in the next chapter apply especially to curve types A, B, C, and D, which have a pronounced central tendency; types E and F cannot be summarized so easily.

SUMMARY

Data that are classified by qualitative characteristics, or attributes, may be summarized and compared by means of *ratios*. On the other hand, the values of a variable that are classified by size at a given point of time are grouped in a *frequency distribution* to facilitate presentation and analysis.

A statistical *ratio* is the quotient of two related values. The base, or denominator, is chosen as the standard with which the numerator is compared, and should be directly comparable with it.

Ratios should be refined, if possible, by adjusting the numerator or denominator to eliminate any extraneous factors obscuring their relationship. The base may be expressed in any convenient multiple of ten units, although the percent form is most common.

Ratios must be interpreted with care, particularly in distinguishing percent change from the difference between two percents. Ratios in tables should be accompanied by the original data to aid in checking figures and in making other comparisons.

In constructing a *frequency distribution*, the range of the variable is divided into intervals, and only the number of values of X in each class is shown, thus sacrificing some detail for conciseness.

The values of X are first *arrayed* by listing them individually or marking them on a tally sheet in the order of their size. The figures are then *grouped* into from 6 to 15 classes so as to show the important characteristics of the data, but without undue detail. Class limits are chosen so that points of concentration, if any, are at midpoints or symmetrical about such points, in order that each midpoint will approximate the average value of X in the class interval. The intervals should be equal in size, if possible. The limits of the classes must be specified unambiguously. Frequencies may be expressed as percents of the total number to facilitate comparisons or to make inferences from samples.

Frequency distributions may be charted by plotting frequencies on

the Y scale above the class midpoints on the X axis. Either a *histogram* (bar chart) or a *frequency polygon* (line chart) may be used. Two frequency distributions may be conveniently compared by plotting the percent frequencies as polygons on the same scales. Frequencies may also be added up from either end and plotted as a cumulative frequency curve or *ogive* to show the number or proportion of values less than or greater than a given amount.

A smooth curve drawn through a histogram or frequency polygon of a continuous distribution approximates the frequency curve for the population from which the sample was drawn, provided the sample is carefully selected and the data do not cluster at certain points.

Frequency distributions may assume a *normal* bell-shaped curve or some other symmetrical form; they may be skewed or asymmetrical either to the left or right; or in extreme cases, they may assume the shape of a reverse J or U.

PROBLEMS

- Given the following information concerning federal credit unions:

AREA	NUMBER OF ASSOCIATIONS	MEMBERS (Thousands)	LOANS MADE DURING YEAR	
			Number (Thousands)	Amount (Millions)
United States.....	8,350	4,502	3,300	\$1,580
Pennsylvania.....	843	433	300	129

- Compute whatever ratios you consider necessary to analyze these data.
 - Write a statement of your findings.
- The American Appraisal Company index of construction costs in 1960 was 722 percent of the 1913 base, and in 1965 was 824 percent of the same base. What is
 - The difference between the 1960 and 1965 figures in percentage points?
 - The percent relation between costs in 1960 and 1965?
 - The percent change from 1960 to 1965?
 - Given the following:

Month	Apparel Sales	Number of Days Store Was Open
February.....	\$31,872	23
March.....	33,084	26

Find the percent change in average daily sales from February to March.

4. The following is quoted from the report of an oil well servicing company to the stockholders: "Foreign operations [in 1965] including export sales, accounted for 15% of consolidated revenue, up from 12% in 1964; and net income was even higher in proportion, one reason being that the majority of the countries have less confiscatory income tax laws than the United States." What additional data would be needed in order to determine the importance of this report?
5. What refinement would you recommend in the denominator of each of these ratios?
 - a) Employees killed in airplane accidents to total number of employees of airlines.
 - b) The number employed in a community to the number of persons in the community.
 - c) The number of Plymouth automobiles manufactured to the total number of motor vehicles sold in the United States.
6. Define and give the purpose of (a) an array, (b) relative frequency distribution, (c) frequency polygon, (d) ogive, and (e) normal curve.
7. Indicate which of the following are correct statements and amend any that are incorrect:
 - a) Points of concentration are always present in an array and should be considered in preparing a frequency distribution.
 - b) All frequency distributions should have at most 14 class intervals.
 - c) Class intervals of unequal width should never be used.
 - d) Class limits should be established so that the average value of the items in each interval is approximately equal to the midpoint of the interval.
 - e) In presenting a distribution of continuous data, the best way to designate the classes is by listing the class midpoints.
8. State wherein each of the following meets or fails to meet the principles of constructing a frequency distribution.

(a)		(b)	
Income	Average Monthly Rent	Age (Years)	Thousands of Persons
Under \$2,000	\$62.70	All ages	5,390
\$2,000-\$2,900	65.40	Under 4	335
\$2,900-\$4,000	70.00	Under 2	87
\$4,000-\$4,900	81.10	4-9	602
\$5,000-\$6,500	93.50	10-15	721
etc.		16-25	1,358
		26-35	1,483
		etc.	

- 9-11. A survey of typical starting salaries offered college men with bachelors' degrees by 200 companies in 1965 showed the following results:

STARTING SALARY*	FIELD				
	Accounting	Sales Marketing	General Business Admin.	Prod. Mgt.	Economics- Finance
425 and under 450.....	2	4	4	1	2
450 " " 475.....	3	7	12	1	3
475 " " 500.....	12	17	15	4	7
500 " " 525.....	16	21	18	9	6
525 " " 550.....	35	16	22	2	7
550 " " 575.....	26	7	12	5	1
575 " " 600.....	8	7	7	2	3
600 " " 625.....	1	2		3	
625 " " 650.....				2	
Number of com- panies reporting.....	103	81	90	29	29

* Class limits in the end classes have been modified slightly in order to facilitate analysis.

NOTE: These data will be used also in Chapters 5 and 6.

SOURCE: Frank S. Endicott, *Trends in Employment of College and University Graduates in Business and Industry* (Evanston, Ill.: Northwestern University Press, 1965), p. 5.

9. a) Plot histograms for two fields in the above table as assigned, using separate graphs.
b) Plot frequency polygons for the same two fields, using either one or two graphs.
c) Compare the merits of the histogram and the polygon in this case.
10. a) Compute a percent frequency table for the two fields assigned in 9(a) above. Use these computations to construct two percent frequency polygons on the same graph.
b) What is the reason for using percent frequencies in comparing two distributions?
c) In what situation would percent frequencies be unnecessary for comparing two distributions?
11. a) Construct a "more than" cumulative frequency table and ogive for one of the fields in the above table as assigned.
b) Construct a "less than" table and ogive for the same field.
c) How many companies offered starting salaries to college men in this field of \$500 and more; of \$550 and more?
d) How many companies offered starting salaries to college men in this field of less than \$575; of less than \$525?
12. a) Make a frequency table, using the 112 items in the four columns assigned to you from the following table (see numbered assignments below table).
b) Give reasons for your choice of class limits and width of class intervals.
c) Draw a graph showing your frequency distribution.
d) What information concerning earnings of women in this plant can be derived from your table and graph?

NOTE: This problem will be continued in Chapters 5 and 6.

DAILY EARNINGS OF 168 WOMEN IN AN ELECTRONIC ASSEMBLY PLANT

(In Dollars)

(a)	(b)	(c)	(d)	(e)	(f)
15.20	18.00	11.20	16.00	20.00	13.60
11.60	14.00	12.00	11.30	12.20	12.00
8.00	12.00	17.60	15.60	8.50	8.00
12.80	12.80	9.50	12.00	14.50	10.00
14.00	11.80	12.00	10.60	16.00	12.60
6.40	9.20	14.00	12.00	12.60	14.00
12.00	7.60	12.00	15.00	12.00	6.50
12.40	14.80	8.20	6.00	8.00	16.00
24.00	18.00	28.00	8.00	19.00	14.00
14.60	16.80	16.80	16.00	22.00	14.60
9.00	14.20	14.40	17.20	15.20	19.20
16.50	12.00	21.20	14.40	10.00	12.30
20.00	12.00	20.00	12.50	14.00	11.60
18.00	21.00	23.00	20.00	16.00	16.40
14.10	8.00	14.00	18.80	16.40	16.00
22.50	16.00	16.10	12.00	12.00	20.00
12.00	24.00	19.90	12.00	23.80	21.40
20.80	19.60	12.90	8.40	28.40	24.00
16.00	27.00	24.00	23.50	17.30	28.80
18.00	20.00	16.00	20.00	18.00	15.20
7.20	10.40	8.00	21.60	14.00	25.00
14.00	15.50	11.80	24.40	11.40	12.00
26.00	21.80	15.00	14.00	24.50	20.40
16.00	14.00	16.00	16.20	6.00	17.60
16.00	6.00	12.40	28.00	20.00	8.80
12.00	16.00	18.40	16.90	16.00	16.00
19.40	12.40	15.50	13.00	12.00	18.00
10.00	16.00	6.00	14.00	13.20	12.00

Assignments:

No.	Columns	No.	Columns	No.	Columns
1.....	a b c d	6.....	a b e f	11.....	b c d e
2.....	a b c e	7.....	a c d e	12.....	b c d f
3.....	a b c f	8.....	a c d f	13.....	b c e f
4.....	a b d e	9.....	a c e f	14.....	b d e f
5.....	a b d f	10.....	a d e f	15.....	c d e f

13. U.S. family personal incomes in 1962 were distributed as follows, according to the *Survey of Current Business* (April 1964):

Income	Percent	Income	Percent
Under \$2,000.....	6.9	\$ 6,000-\$ 7,499.....	16.0
\$2,000-\$2,999.....	6.2	\$ 7,500-\$ 9,999.....	18.6
\$3,000-\$3,999.....	8.2	\$10,000-\$14,999.....	14.8
\$4,000-\$4,999.....	9.8	\$15,000 and over.....	8.7
\$5,000-\$5,999.....	10.8	Total families.....	100.0

- a) Criticize the choice of class intervals and class limits.
b) Plot a histogram of this distribution. Then draw a smooth curve to approximate the true continuous distribution of incomes. What type of frequency curve is this—normal, negatively skewed, etc.?

14. An automobile advertisement lists the following distribution of gas mileage reported by owners of its new cars:

<i>Miles per Gallon</i>	<i>Percent</i>	<i>Miles per Gallon</i>	<i>Percent</i>
15 and under 16*	6	19 and under 20	14
16 and under 17	10	20 and under 21	18
17 and under 18	16	21 and under 22*	12
18 and under 19	24	Total owners	100

*Open-end classes have been assigned arbitrary limits to facilitate later computations.

- a) Plot a histogram of gas mileage, and draw a smooth curve through it to iron out sampling irregularities and approximate the continuous distribution of mileage performance for the whole population of car owners. What type of frequency distribution is this?
- b) List a cumulative frequency distribution and draw an ogive showing the percent of owners reporting a given gas mileage or more. From this curve, half the owners get what gas mileage or more? The most economical fourth of the owners get what gas mileage or more? (Give results to nearest tenth of a gallon.)
15. You are comparing two brands of a certain type of electron tube. You obtain the following frequency distributions for their life in hours.

DISTRIBUTION OF LIFE OF ELECTRON TUBES

BRAND A AND BRAND B

Life (Hours)	Frequency		Relative Frequency, Percent	
	Brand A	Brand B	Brand A	Brand B
Under 50	1	3	0.8	3.8
50 and under 100	8	8	6.7	10.0
100 and under 150	18	12	15.0	15.0
150 and under 200	40	14	33.3	17.5
200 and under 250	26	13	21.7	16.3
250 and under 300	12	10	10.0	12.5
300 and under 350	6	9	5.0	11.2
350 and under 400	3	6	2.5	7.5
400 and under 450	2	3	1.7	3.8
450 and under 500	1	1	0.8	1.2
500 and above	3*	1*	2.5	1.2
Total	120	80	100.0	100.0

* The mean life for those tubes still burning after 500 hours was 700 for Brand A and 600 for Brand B.

SOURCE: Company records.

- a) Plot on the same chart the relative frequencies of the two brands. (For this purpose, omit the class 500 and above.) Why should you use percentages rather than the actual number of tubes?
- b) Are these frequency distributions fairly normal, skewed to the left, skewed to the right, J-shaped, or U-shaped?

- c)* Use your chart to compare the two frequency distributions.
- d)* Calculate cumulative frequency distributions for the two brands of tubes. Then plot these distributions on a chart. At what life are approximately 50 percent of Brand A tubes still burning? 50 percent of Brand B tubes? (This can be obtained from your chart—where the cumulative frequency curves cross the 50 percent cumulative frequency line.) Using this result and your analysis in part (*c*) above, which tube do you think you should buy to obtain greater total life? Why?
- e)* Suppose your company had a policy of replacing all tubes after 150 hours. Would this change your answer to (*d*) above?

SELECTED READINGS

Selected readings for this chapter are included in the list that appears on page 139.

5. AVERAGES

A BASIC purpose of statistical analysis is to develop concise summary figures that will describe unwieldy masses of raw data. The initial stages in this analytic process have already been described—that is, appraising the accuracy of data, classifying facts for tabulation and graphic presentation, and condensing a long list of separate values into a frequency distribution.

An important type of summary measure needed in statistical analysis is the *average*.¹ Averages are familiar to everyone in such examples as average weekly wages, average prices of securities, a man of average income, a medium-sized house, and the usual rate of interest charged a bank's customers. Careful analysis of these examples shows that they involve several different concepts of "average" which should be distinguished from each other. No single average can be used indiscriminately.

The most common averages are (1) the arithmetic mean, (2) the median, and (3) the mode. The first is determined by calculation, the second by its position in an array, and the third by finding the point about which values of the variable cluster most closely. These will be described in turn. Other calculated averages, such as the modified mean and the geometric mean, have important special uses but will not be emphasized in this chapter.

THE ARITHMETIC MEAN

The most common average is the arithmetic mean or, more simply, the mean. The term "average," when used alone, usually refers to the

¹ An average is sometimes called a "measure of central tendency" because individual values of the variable usually cluster around it. Averages are useful, however, for certain types of data in which there is little or no central tendency.

mean. The *mean* of any series of values is found by adding them and dividing their sum by the number of values. In terms of symbols to be used in this chapter, the mean of n values of a variable X is calculated by adding X values and dividing the sum by n .

Ungrouped Data

The general method of computing the mean is the same whether the data are ungrouped or grouped in a frequency distribution, but the formulas look a little different. As an example of ungrouped data, consider a man working at piece rates who earns \$2.80, \$3.05, \$3.00, and \$3.15 in four successive hours. His mean hourly earnings is found by adding his earnings for the four hours and dividing by 4. The earnings total \$12.00, so the mean is \$3.00. This process is generalized by the following formula:

$$\bar{X} = \frac{\Sigma X}{n}$$

where \bar{X} (read "X bar") is the symbol for the mean of the variable X (hourly earnings in dollars); Σ is the Greek letter capital sigma (corresponding to our S), which means "the sum of"; and n is the number of values.²

When a variable has a number of identical values, multiplication can be used as a short-cut for addition in totaling X . Thus, to find the average dimension of the 63 gears in Table 4-2, one could add the 63 figures in panel A, but it would be easier to multiply each dimension in panel B by its *frequency* and add the products as follows: $1(.4270) + 4(.4265) + 10(.4260) + \dots$. Specifically, since there are ten gears measuring .4260, it is simpler to multiply 10 by .4260 than to add .4260 ten times. The whole process is summarized by the formula

$$\bar{X} = \frac{\Sigma fX}{n}$$

where f is the symbol for frequency, and ΣfX means that each different value of X is multiplied by its frequency and the products (fX) are then added. Using either formula,

$$\bar{X} = \frac{26.7820}{63} = .4251, \text{ the mean dimension in inches}$$

² Strictly speaking, the symbols \bar{X} and n apply only to sample data. In later chapters, μ (the Greek letter mu) will be used to designate the mean of an entire population and N , the number of values in the population. Hence, $\mu = \Sigma X/N$.

The Weighted Mean. In many types of problems, the values to be averaged are of different degrees of importance. In such cases, each value is multiplied by a numerical weight based on its relative importance, and the total is divided by the sum of the weights. The result is called a *weighted mean*. The weights are handled just as if they were frequencies. Hence, a weighted mean can be computed by the above formula—taking f as the weight and n as the sum of the weights.

Thus, an aptitude score may be based on an English test with weight 2 and a mathematics test with weight 1. The weights total 3. If a person makes 90 and 60, respectively, on these tests his combined aptitude score is

$$\bar{X} = \frac{\Sigma fX}{n} = \frac{2(90) + 1(60)}{3} = \frac{240}{3} = 80$$

Weighted means are used extensively in the construction of index numbers, to be described in Chapter 18.

All means can be regarded as weighted in some way, either explicitly or implicitly. From this point of view, the “unweighted” mean is one in which the weights are all equal. In computing any mean, therefore, it is important to use appropriate weights. In averaging the ratios of profits to sales for 30 retail grocers, for example, the total profits for all 30 grocers can be divided by their total sales to allow the larger firms more weight in the results, or the firms may be weighted equally by taking a simple average of the 30 ratios. If the larger grocery stores are more profitable than the smaller ones, the weighted mean profits-to-sales figure will exceed the unweighted mean.

Grouped Data

The mean of data grouped in a frequency distribution is computed in the same way as described above. In a frequency distribution, however, the *midpoint* of each interval is used to represent all values of X in the interval. Accordingly, each midpoint is multiplied by the number of values in that class. The sum of these products is then divided by the total number of values of X to find the mean.

The formula for computing the arithmetic mean from a frequency distribution is therefore

$$\bar{X} = \frac{\Sigma fX}{n}$$

where

- \bar{X} = the arithmetic mean computed from a frequency distribution;
 X = the midpoint of each interval;
 f = the frequency (number of values of X) in that interval;
 fX = their product;
 ΣfX = the sum of these products; and
 n = the total number of values or the sum of the frequencies.

In calculating the arithmetic mean for the earnings of machine tool operators shown in Table 5-1, the midpoint of each interval is used to

Table 5-1
 DIRECT METHOD OF COMPUTING THE ARITHMETIC MEAN
 FROM A FREQUENCY DISTRIBUTION
 HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL OPERATORS

Hourly Earnings, Dollars	(1) Class Midpoint X	(2) Number of Operators, Frequency f	(3) Frequency \times Midpoint fX
2.25 and under 2.35.....	2.30	2	4.60
2.35 and under 2.45.....	2.40	23	55.20
2.45 and under 2.55.....	2.50	49	122.50
2.55 and under 2.65.....	2.60	63	163.80
2.65 and under 2.75.....	2.70	45	121.50
2.75 and under 2.85.....	2.80	25	70.00
2.85 and under 2.95.....	2.90	3	8.70
2.95 and under 3.05.....	3.00	4	12.00
Total.....		214	558.30

SOURCE: Table 4-4.

represent all earnings figures in that interval. The total earnings for the two operators in the first class are thus computed to be $2.30 \times 2 = 4.60$. Applying this procedure to the other classes yields the products listed in column 3, for which the grand total is 558.30. Then dividing this total by 214, the number of operators, the arithmetic mean is found to be \$2.609 per hour. That is,

$$\begin{aligned}
 \bar{X} &= \frac{\Sigma fX}{n} = \frac{558.30}{214} \\
 &= 2.609
 \end{aligned}$$

The mean computed from a frequency distribution is subject to a slight *error of grouping*, since all values are rounded off to the nearest class midpoint. This error would be nil if the mean of the values in each

class were equal to the midpoint, or if the plus and minus errors of grouping in the various classes offset each other. The error can be minimized by placing the midpoints of class intervals at points around which the data tend to cluster or midway between such points within intervals. Grouping errors of opposite sign often tend to offset each other, so that the grouped mean is usually very close to the ungrouped mean, particularly if the number of values is large and the distribution is nearly symmetrical. Thus, the arithmetic mean of \$2.609 per hour obtained from the frequency distribution is only \$.003 greater than the exact mean of \$2.606 per hour computed from the original figures.

The arithmetic mean and other statistical measures are often computed from a frequency distribution rather than from ungrouped data despite minor errors of grouping because (1) it is much easier to calculate the mean from grouped data when the number of original values is large and (2) many types of data are available only in the form of frequency distributions.

Short-Cut Methods. When computing the mean from a frequency distribution, short-cut techniques are available that will reduce the amount and difficulty of the necessary calculations. One such method will be treated in detail in the following chapter, in conjunction with a short method for computing the standard deviation.

Open-End Distributions. On some occasions it is necessary to compute the mean from a frequency distribution having open-end classes whose lower or upper limit is not indicated, such as a salary class "\$425 or less." Although open-end intervals should be avoided ordinarily, it is possible to compute the mean from open-end distributions provided either the individual values, their average, or their total is available for each open-end class to supply the missing data. Simply use the average of the open-end interval as the X or midpoint value for that interval in the computation of the overall arithmetic mean. If the mean or total of the open-end interval is missing, the mean can be computed only by guessing at these values. In such instances the median, modified mean, or mode should be used in preference to the mean, since they do not depend on extreme values.

Attribute Data

When the data for analysis are attributes (i.e., classified into only two categories), the arithmetic mean has a special interpretation. A ratio or proportion may be considered to be a special case of the arithmetic mean in which all the values are ones or zeros. Thus, if 20 out of 100 bolts inspected are defective, and we count the defectives as ones and the

others as zeros, the *average* of the 20 ones and the 80 zeros is 0.20, which is the same as the *proportion* defective.

THE MEDIAN

The median of any set of data is the middle value in order of size if n is odd, or the mean of the two middle items if n is even. When there are a few very large or small values, the median is often superior to the mean as an average. For example, the *Monthly Labor Review* reports median wages and salaries by occupations, and *Dun's Review and Modern Industry* reports median operating ratios for small samples of business firms because the median represents the typical middle man or firm undistorted by large values that so greatly affect the mean. To cite a specific case, the median income of American families and unattached individuals in 1963 was \$6,140, whereas the mean was \$7,510, according to the *Survey of Current Business* for April 1964.

The median can sometimes be found when other averages are not defined because individuals are not measured quantitatively. For example, employees in a plant can be rated by arranging them in order of merit without assigning a numerical grade to each individual. To find the value of the median under these conditions, only one or two individuals need be measured or graded. The median can also be computed in an open-end frequency distribution, while the mean cannot, if the end values are unknown.

Ungrouped Data

In ungrouped data, the median is most easily found when the values are arranged in an array. Consider the price-earnings ratios 19.6, 17.3, 19.2, 14.0, and 29.9 (i.e., common stock prices divided by earnings per share) for five electronics companies. Arranged in order of size, the five ratios are

14.0, 17.3, 19.2, 19.6, and 29.9

The median is then the middle value, or 19.2. If a sixth ratio, 30.0, were added, the median would be the mean of the two middle items 19.2 and 19.6, or 19.4. In general, the median in an array is not computed from a formula but is selected as the value whose rank or "order number" is $n/2 + 1/2$, counting from the lowest value. Thus, for the six ratios above, the order number of the median is $6/2 + 1/2 = 3\frac{1}{2}$, i.e., halfway between the third and fourth values.

This example illustrates an important advantage of the median over the mean. The ratio of the price of a stock to the earnings per share is

sometimes very large when the earnings are abnormally small, as in the case of the 29.9 ratio above. Because of this figure, the mean (20.0) exceeds any of the other four ratios. The median is often more reliable than the mean in samples from populations in which such extreme deviations occur, because the reliability of the mean is greatly affected by extreme deviations, while the reliability of the median depends chiefly upon the degree of clustering about the median of the population.

Grouped Data

When data are grouped in a frequency distribution, the median falls in the class interval whose frequency is the first to make the cumulative frequency greater than $n/2$. It is convenient to call this the median class. The median may then be located within the median class by means of the interpolation formula

$$Md = L + \frac{i(n/2 - F)}{f}$$

where

Md = the median;

L = the lower limit of the median class;

i = the width of the median class;

f = the frequency for the median class;

F = the cumulative frequency for all classes below the median class;

n = the total number of values of X (the sum of all frequencies).

In applying this formula to the earnings data of Table 5-1 above, the first step is to locate the class that contains the middle value, i.e., the one ranked $n/2 = 214/2 = 107$.³ By cumulating the f column, the successive subtotals are found to be 2, 25, 74, 137, etc. The first subtotal to exceed $n/2$ is 137. Accordingly, the *fourth* class is the median class. Its lower limit is $L = 2.55$; its frequency is $f = 63$; the cumulative frequency for X less than L is $F = 74$; and the interval is $i = 0.10$. Substituting these values in the formula, the median is:

$$\begin{aligned} Md &= L + \frac{i(n/2 - F)}{f} \\ &= 2.55 + \frac{.10(107 - 74)}{63} \\ &= 2.55 + .052 \\ &= 2.602, \text{ or } \$2.602 \text{ per hour} \end{aligned}$$

³ The middle value interpolated over a continuous range is at the exact midpoint $n/2$ in rank, rather than $n/2 + 1/2$ as in discrete data.

This value is only an approximation to the median of the original ungrouped data, since it is interpolated on the assumption that values of X in the median class are *evenly distributed* over that interval. In this case the true median, taken from the original data in Table 4-3, is exactly \$2.60, because the earnings around the median cluster at this point.

About half of the 214 earnings are smaller than the median of \$2.60 and about half are larger. The proportion on each side of the median is

Table 5-2
FAMILY INCOMES IN THE NORTHEASTERN
STATES, 1964

Income	Percent of Families	Cumulative Percentage
Under \$3,000.....	12	12
\$ 3,000-\$ 4,999.....	15	27
5,000- 6,999.....	21	48
7,000- 9,999.....	25	73
10,000- 14,999.....	19	92
15,000 and over.....	8	100
Total.....	100	

NOTE: Excludes unrelated individuals
SOURCE: U.S. Department of Commerce, *Consumer Income*, Current Population Reports, Series P-60, no. 47, September 24, 1965, p. 4.

exactly one half when the median is between the two middle values. In fact, a vertical line at the interpolated median always divides the histogram into two parts whose areas are equal. Nevertheless, the proportion of items on each side of the median is sometimes more or less than one half. In ungrouped data, one or more values may be equal to the median so that the proportion of values smaller (or greater) than the median may be considerably less than one half—it can never be greater. In grouped data, more than one half of the original values may be on one side of the interpolated median because of uneven distribution of values within the median class. For these reasons, it is better to say that the proportion of values on each side of the median is only *approximately* equal to one half.

Open-End Distributions. Since the median is not affected by the size of extreme values, it can be determined in an open-end distribution

$$\begin{aligned}
 Md &= L + \frac{i(n/2 - F)}{f} \\
 &= 7,000 + \frac{3000(50 - 48)}{25} \\
 &= 7,240 \text{ or } \$7,240 \text{ income}
 \end{aligned}$$

such as that of family incomes in the Northeast, presented in Table 5-2. The percent figures here can be treated as ordinary frequencies, and the median is found to be \$7,240. This indicates that in 1964 about half of the families received over \$7,240, and about half received less.

Graphic Interpolation. The median in a frequency distribution can be obtained graphically from a cumulative frequency curve or ogive. For example, the median hourly earnings of the 214 machine tool operators can be found from either ogive in Chart 4-4. A horizontal line is drawn from the 50 percent ordinate on the right vertical scale (107 or $n/2$ on the left scale) until it intersects the ogive. (The two ogives in Chart 4-4 intersect at the same point.) The X value of this point, which is read as \$2.60 on the bottom scale, is the median. The graphic method yields the same result as the interpolation formula of the preceding section, except for errors in plotting and reading the scale.

THE MODIFIED MEAN

A modified mean is the mean of a central group of values in an array or frequency distribution, omitting any very large and small values that are so extreme and atypical as to distort the overall mean. Indeterminate items in open-end classes may also be omitted. The analyst must use his judgment as to how many values to discard. Usually the same predetermined number of items is omitted at each end of the array or distribution, as in seasonal analysis (described in Chapter 20), but there are many variations. The National Bureau of Economic Research in averaging business cycles omits certain extreme values that are judged to be erratic, but does not exclude indiscriminately any fixed number of items at both ends of an array.⁴

As more and more end items are omitted until only the middle one or two are left, the modified mean becomes the median. Thus, there is a whole family of modified means, of which the mean itself includes the maximum and the median the minimum number of central values in an array. The intermediate means are, therefore, compromises between the mean and the median, selected to combine the best features of both.

THE MODE

The *mode* in statistics means just what it does in the dictionary—the prevalent or most frequently encountered thing. More precisely, the mode is defined as the *value which occurs most often* or the value around which there is the *greatest degree of clustering*. The modal wage

⁴ Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946), p. 496.

is the one received by the greatest number of workers. The modal interest rate for bonds is the one that occurs more often than any other. If the most common or usual value is the one needed for a business decision, the *mode* is the appropriate type of average to use.

It is particularly important that the data used to determine the mode be homogeneous or enough alike to be comparable. Wage data that include skilled and unskilled workers, men and women, or industrial and farm workers may be so diverse that the modal wage would have little meaning. Such data might also have two or more modes of about equally high frequency. The mode is ordinarily meaningful only if there is a marked concentration of values about a single point.

Ungrouped Data

The mode can occasionally be determined directly from ungrouped data. When a large proportion of values are equal, no process of grouping could dislodge this value from its modal position. This is especially true of discrete data having only a limited number of possible distinct values. For example, if a bank charges the general run of its customers 6 percent interest on commercial loans, then 6 percent is the mode of interest rates, irrespective of what rates apply in special cases. Similarly, a survey indicates that more parents prefer to have three children than any other number. Thus, three is the modal family size preferred by parents.

Grouped Data

Most types of data, however, must be grouped in a frequency distribution in order to locate the mode. To illustrate, in the array of hourly earnings listed by cents in Table 4-3, the most frequently occurring rate is \$2.63, but \$2.70 is almost as popular; and there are other scattered points of concentration, such as \$2.50 and \$2.75, which cause doubts as to where the major area of concentration really is. By grouping the earnings as in Table 5-1, however, there appears only a single mode. This occurs in the \$2.55 to \$2.65 interval. The modal interval can be described by saying, "More earnings fall in the \$2.55 to \$2.65 class than in any other."

The value of the mode within this interval may be estimated graphically in a continuous distribution by drawing a smooth curve through the histogram so that the area cut from each bar is about equal to the area added to that bar by the curve. The mode is then the *X* value at the peak of the frequency curve. Thus, in Chart 4-5 the modal price of laying mash is about \$4.57 per hundredweight.

Interpolation formulas are also used to locate a "single-valued" mode within the modal interval.⁵ More simply, the midpoint of the modal interval could be taken as the mode, but this is recommended only if values cluster at this point. Ordinarily, a single-valued estimate of the mode is neither accurate nor necessary in practice. In the relatively rare cases in which the mode is needed, it is usually enough to cite the modal interval.

The modal interval itself is only a rough estimate, since it depends on the choice of class limits. Grouping the data in different class intervals will produce different values of the modal interval. In some types of data, therefore, the mode is practically indeterminate. Hence, the mode or modal interval should be used only if the problem specifically requires the most usual or common value as an average rather than the middle or the mean value.

THE GEOMETRIC MEAN

The geometric mean is sometimes appropriate for averaging index numbers, percentages, and other ratios. It may also be a good type of average for frequency distributions of absolute data that are skewed to the right (see Chart 4-6D), provided the distribution of logarithms is more nearly symmetrical. Moderately symmetrical distributions of absolute values with only a few extreme items, however, can best be averaged by a median rather than by the geometric mean.

The geometric mean has certain disadvantages that have limited its use. It is difficult to compute and to interpret. Hence, the arithmetic mean is actually used for computing index numbers (Chapter 18) and other averages of ratios for which the geometric mean might seem more appropriate. Also, the geometric mean cannot be computed if any of the values is zero or negative. Profit and loss data, for example, could not be averaged in this manner.

The geometric mean is computed in exactly the same way as the arithmetic mean, except that the logarithms of the numbers are averaged to find the logarithm of the geometric mean. The geometric mean of X thus may be defined as the antilogarithm of the arithmetic mean of

⁵ See Spurr, Kellogg, and Smith, *Business and Economic Statistics* (1st ed., Homewood, Ill.: Richard D. Irwin, Inc., 1954), pp. 208-10, for a description of the most common method. The mode may also be estimated from the mean and median, as follows:

$$\text{Mode} = \text{mean} - 3(\text{mean} - \text{median})$$

This formula is based on the tendency of the median to fall roughly one third of the way from the mean toward the mode in a continuous distribution of only moderate skewness. Unfortunately, frequency distributions of economic data are seldom smooth enough to justify the use of this formula in estimating the mode.

$\log X$.⁶ Therefore, except that $\log X$ is used in place of X , the main formulas used to find the geometric mean are the same as the corresponding ones for the arithmetic mean.

Ungrouped Data

For ungrouped data,

$$\log G = \frac{\sum \log X}{n}$$

where G is the geometric mean.

The geometric mean of the price-earnings ratios for five electronics stocks is computed in Table 5-3.

Table 5-3

GEOMETRIC MEAN OF FIVE PRICE-EARNINGS RATIOS

Common Stock	Price-Earnings Ratio (X)	Logarithm of Price-Earnings Ratio (log X)
A.....	19.6	1.2923
B.....	17.3	1.2380
C.....	19.2	1.2833
D.....	14.0	1.1461
E.....	29.9	1.4757
Total.....	100.0	6.4354

Substituting in the formula:

$$\begin{aligned} \log G &= \frac{\sum \log X}{n} & G &= \text{antilog} (\log G) \\ &= \frac{6.4354}{5} & &= \text{antilog } 1.2871 \\ &= 1.2871 & &= 19.4 \end{aligned}$$

For comparison, the arithmetic mean of these five ratios is 20.0. The geometric mean is always less than the arithmetic mean for a series of different values.

Grouped Data

The geometric mean may be similarly computed from a frequency distribution by multiplying the logarithm of each class midpoint, $\log X$, by the class frequency f before averaging the results. That is,

⁶The geometric mean may also be defined as the n th root of the product of n values ($G = \sqrt[n]{X_1 \cdot X_2 \cdot \dots \cdot X_n}$), but this form is not popular because one would usually find the results by logarithms anyway, and this approach leads to that explained in the text.

$$\log G = \frac{\sum f \log X}{n}$$

This formula will not be illustrated here, since its practical use is somewhat limited.

WHICH AVERAGE TO USE?

Much of the chapter thus far has been devoted to methods of computing the various types of averages. In the course of the several explanations, the distinctive features of the measures have been set forth in some detail but in incidental fashion. At this point, the reader may well ask, "Which of these various averages should I use?" or "When ought I to use one or the other of the averages described?"

There is no arbitrary single answer that can be given to these questions. The selection of the proper average depends upon three main factors:

1. The concept of the typical value required by the problem. Is a composite average of all absolute or relative values needed (arithmetic or geometric mean) or is a middle value wanted (median) or the most common value (mode)?
2. The type of data available. Are they badly skewed (avoid the mean), gappy around the middle (avoid the median), or lacking a major point of concentration (avoid the mode)? In particular, the choice between the arithmetic mean and median of a sample depends on the shape of the frequency curve for the population. Refer to Chart 4-6. If the distribution is normal (panel A) or flat-topped with few extreme values (panel B, lower curve), the mean has a smaller sampling error than the median. That is, the mean of the sample is likely to be closer to the true mean of the population than the median of the sample is to the true median. On the other hand, if the distribution is sharply peaked around the median and includes some extreme values (panel B, higher curve), the median has the smaller sampling error. This is because the clustering around the population median makes the sample median more accurate, and extreme values make the sample mean erratic.
3. The peculiarities or characteristics of the averages themselves. These will be summarized below, under "Characteristics of Averages."

As a rule of thumb, the arithmetic mean should ordinarily be used as a simple, widely understood average which gives due weight to all

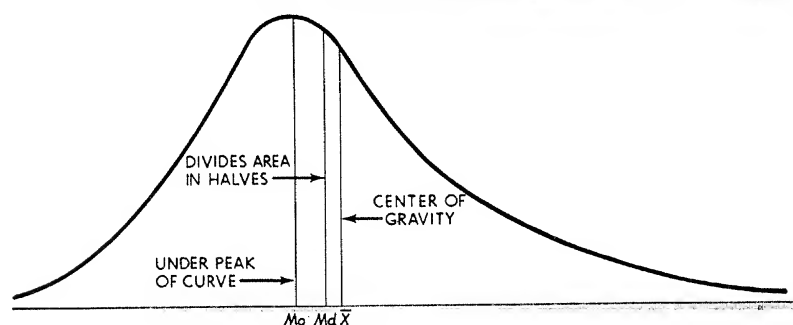
values. If the items are few in number or erratic in value, a modified mean is desirable. The median is commonly preferred to the mean if a simpler, middle value is needed—particularly if the data are badly skewed, as is common in economic measurements. Finally, the mode may be used if the most usual or common value is wanted.

CHARACTERISTICS OF AVERAGES

The arithmetic mean, median, and mode have the same value in a symmetrical “normal” distribution. If the distribution is skewed, the mode remains under the highest point of the curve, the arithmetic mean is pulled out in the direction of the extreme values, and the median, which is affected by the *number* of extreme items but not their *value*,

Chart 5-1

RELATIONSHIP OF ARITHMETIC MEAN, MEDIAN, AND MODE
IN A POSITIVELY SKEWED DISTRIBUTION



tends to fall between the mean and the mode. The mean, median, and mode thus rank in the order given. The geometric mean, which gives less weight to large absolute values, is smaller than the arithmetic mean in either case.

Chart 5-1 shows the relation of the arithmetic mean, median, and mode in a positively skewed distribution—by far the most common type in business and economic data. Here the arithmetic mean is the largest value, and the mode is the smallest. Thus, family incomes in 1963 had a mean value of \$7,510 and a median of \$6,140, as cited above, but the mode was only \$5,210. The mean is the X value of the center of gravity. That is, if the area under the curve were a solid piece of metal, a fulcrum under \bar{X} would balance it. The median divides the *area* under the curve (i.e., the total frequency) into two equal parts. The mode is the value of X under the highest point of the curve.

The characteristics of the individual averages are listed below.

Arithmetic Mean

1. The arithmetic mean is the most widely known and widely used average.
2. It is, nevertheless, an artificial concept, since it may not coincide with any actual value.
3. It is affected by the value of every item, but
4. It may be affected too much by extreme values.
5. It can be computed from the original data without forming an array or frequency distribution, or from the total value and number of items alone.
6. Being determined by a rigid formula, it lends itself to subsequent algebraic treatment better than the median or mode.
7. It is less affected by sampling errors than the median, in a normal or flat-topped distribution.

Median

1. The median is a simple concept—easy to understand and easy to compute.
2. It is affected by the number but not the value of extreme items.
3. It is widely used in skewed distributions where the arithmetic mean would be distorted by extreme values.
4. It may be located in an open-end distribution or one where the data may be ranked but not measured quantitatively.
5. It is unreliable if the data do not cluster at the center of the distribution.
6. The median will have a smaller sampling error than the mean if the data *do* cluster markedly at the middle or if there are abnormally large or small values. Such sharply peaked and long-tailed distributions are fairly common in economic data.

Modified Means

1. Modified means are compromises between the arithmetic mean and the median, so they combine the characteristics of both.
2. Any one of several modified means may be used, depending on the number of items selected by the analyst.
3. Modified means are particularly adapted to a small or erratic group of values in which neither the mean nor the median is satisfactory.

Mode

1. The mode can best be computed from a frequency distribution, unless one value predominates in an array.
-

2. It can be located in open-end distributions, since it is not affected by either the number or value of items in remote classes.
3. The mode is erratic if there are but few values or zigzag frequencies—particularly if there are several modes or peaks.
4. It is affected by the arbitrary selection of class limits and class intervals.

Geometric Mean

1. The geometric mean averages ratios or percentages in the same way that the arithmetic mean averages absolute values, so it is also characterized by points 2 to 7 under "Arithmetic Mean" above, as applied to the logarithms of numbers.
2. The geometric mean is a difficult concept and hence is not widely understood.
3. It cannot be computed if the series contains zero or negative values.

SUMMARY OF FORMULAS

Since the characteristics of the various averages have been summarized above, the chapter may be concluded by listing the principal formulas used:

Type of Average	Ungrouped Data	Grouped Data
Arithmetic mean...	$\bar{X} = \frac{\Sigma X}{n}$	$\bar{X} = \frac{\Sigma fX}{n}$
Median.....	Value # $n/2 + \frac{1}{2}$ in an array	$Md = L + \frac{i(n/2 - F)}{f}$
Modified mean....	Same as \bar{X} , for central values	
Mode.....	Most common value	Same
Geometric mean...	$\log G = \frac{\Sigma \log X}{n}$	$\log G = \frac{\Sigma f \log X}{n}$

PROBLEMS

1. One method of saving money regularly is to buy common stock at periodic intervals. Is it better policy, then, to buy the same number of shares in a company each year or to invest a constant number of dollars, irrespective of the price of the stock?

To illustrate, Investor A buys 20 shares of Aerojet-General and 20 shares of General Motors common stock at the approximate midyear price, listed below, in each of the years 1961-65. Investor B invests \$1,000, as nearly as possible, in each of these stocks at the same times and prices. His results are detailed in the table. General Motors advanced and Aerojet-General declined in price over this period.

COMMON STOCK PURCHASES BY INVESTOR B

(MIDYEAR PRICES)

	Aerojet-General			General Motors		
	Price per Share	Shares Bought	Total Cost	Price per Share	Shares Bought	Total Cost
1961.....	\$ 79	13	\$1,027	\$ 47	21	\$ 987
1962.....	51	20	1,020	49	20	980
1963.....	54	19	1,026	71	14	994
1964.....	29	34	986	90	11	990
1965.....	27	37	999	97	10	970
Total.....	\$240	123	\$5,058	\$354	76	\$4,921

- Give the average cost per share for Investor A (constant shares) and Investor B (constant dollars), for each stock.
 - Which investor achieved the lower average cost for Aerojet-General? For General Motors?
 - Explain these differences in terms of the weights used in computing the averages.
- In the "dollar-averaging" method of investment, the same amount of money is invested each month in a variable number of shares of common stock. Thus, \$50 will buy one share of a stock selling at \$50 a share in one month, but two shares of that stock if it sells at \$25 in another month. The three shares then cost \$100, or an average of $\$33\frac{1}{3}$ per share, as compared with the average market price of $\$37\frac{1}{2}$ in the two months $[(50 + 25) \div 2]$, irrespective of whether the market is rising or falling. Explain this apparent anomaly in terms of the two types of averages represented.
 - An investor owns three stocks on which he receives the following dividends in 1964 and 1966:

Stock	1964			1966		
	Investment	Dividend	Yield	Investment	Dividend	Yield
A.....	\$ 8,000	\$ 480	6%	\$ 5,000	\$300	6%
B.....	5,000	200	4	12,000	480	4
C.....	6,000	480	8	2,000	160	8
Total.....	\$19,000	\$1,160		\$19,000	\$940	
Average yield.....			6.11%			4.95%

- How are the average yields obtained?
 - Inasmuch as none of the individual yields has changed, how do you explain the decrease in average yield?
- From Chapter 4, Problem 12:
 - Compute the arithmetic mean from your frequency distribution. (Indi-

- cate all computations in this and following problems.) Discuss the grouping errors that affect this value.
 - b) Find the median both from the original data and from your frequency distribution. If these values differ, explain why.
 - c) What does the comparison of the mean and median reveal about the shape of the distribution?
 - d) State the modal interval. Which of the three averages is most meaningful in this case? Why?
5. a) Compute the mean starting salary offered to college men, shown in Chapter 4, the table preceding Problem 9, in whichever of the five fields is assigned.
 - b) Is this mean more or less accurate than one computed from the original ungrouped salary data? Why?
6. a) Find the median starting salary for whichever field was assigned in Problem 5 above.
 - b) Give the modal interval for the same field.
 - c) Explain the difference in the meaning of these two averages.
 - d) If the last four classes had been grouped into one class and labeled "\$550 and over" which measure or measures would have been affected—the mean, median, or mode? Why?
7. The durations of ten business cycles in the United States from 1919 to 1961, measured from trough to trough, were 28, 36, 40, 64, 63, 88, 48, 58, 44, and 34 months, respectively, according to Table 21-1.
 - a) List the mean, median, and all possible modified means of these periods.
 - b) Which of these averages is preferable? Why?
 - c) What is the difficulty in computing the mode for these figures?
8. Under a wages-and-hours law it is considered desirable that the number of hours of work per week should be standardized for some 250 establishments, all now operating under similar conditions except with respect to hours of work. What should be the standardized number of hours (a) if the object is to keep the total hours of work the same and (b) if the object is to change as few establishments as possible?
9. a) Compute the geometric mean for the business cycle data in Problem 7.
 - b) Is this value preferable to the arithmetic mean as a measure of average cycle length? Explain.
10. Regarding the dimensions of 63 gears in Table 4-2:
 - a) Is this distribution discrete or continuous? Symmetrical or skewed to the right or left?
 - b) Find the mean and median to the nearest 0.0001 in. (Express data as deviations from .4250 to simplify calculations.)
 - c) Which type of average is usually the best estimate of the corresponding population value for a distribution of this kind? Why?
11. Chapter 4, Problem 13, reports the distribution of family incomes in 1962. The mean income was stated as \$8,151.

- a) Estimate the median income. What is its significance?
 b) Give the modal interval.
 c) Explain why the mean, median, and mode differ so widely in value.
 Which is the best measure of typical family income? Why?
12. In Chapter 4, Problem 14:
 a) Compute the mean mileage per gallon.
 b) Interpolate to estimate the median mileage.
 c) What does the difference between the mean and median indicate about the skewness of this distribution?
13. Age of XYZ Refrigerators turned in for new models in a recent survey is

Years	No. of Refrigerators
0 and under 1	10
1 " " 2	19
2 " " 3	26
3 " " 4	18
4 " " 5	13
5 " " 6	8
6 " " 7	3
7 and over	3*
Total	100

* The average age of these three refrigerators is $10\frac{1}{2}$ years.

- a) What is the arithmetic mean of the ages of these 100 refrigerators?
 b) Estimate the median age of refrigerators to the nearest year.
14. A trucking concern kept statistics for several years on two makes of tires. It found the following results:

Tire	Median, Miles	Mean, Miles
A.....	25,000	27,000
B.....	27,000	25,000

Assuming that the two tires sell at the same price, which make would you advise the trucking concern to purchase? Why?

15. The U. B. Glad Company operates a small bulk plant which wholesales gasoline to independent retailers. Last week's sales are shown:

Gallons of Gasoline, Thousands	No. of Sales
0 and under 10	10
10 " " 20	20
20 " " 30	30
30 " " 40	25
40 " " 50	15
50 " " 60	10
60 " " 70	5
70 " " 80	5
Total	120

- a) Compute from the above frequency distribution the total number of gallons sold last week.
- b) Compute the average (mean) gallons per sale.
- c) Is the mode above or below 25,000 gallons? How do you know?
- d) Compute the median sale.
16. The president of a company states that the shares of the company are widely distributed. To illustrate his point, he presents the following frequency distribution:

Shares Held	Stockholders, Thousands
1-10	10
11-20	18
21-50	20
51-100	12
101-500	4
501-1,000	2
Above 1,000*	1
	67

*The average number of shares for stockholders in this group is 2,500 shares.

- a. Do you agree with the president's statement? Why?
- b. What is the mean number of shares held? What is the median number of shares held?

SELECTED READINGS

Selected readings for this chapter are included in the list that appears on page 139.

6. DISPERSION

IN THE TWO preceding chapters, attention has been centered on two basic methods of describing a set of data: first, the frequency distribution, which groups a large number of values into a few classes; second, the average, which summarizes the typical value. These devices are useful and important, but they do not describe all of the important characteristics of the figures. Other measures are needed to show how the data vary about the average, because this variation is sometimes as important as the average itself.

There are four important characteristics of a distribution of values which may be described by summary measures:

1. Average—typical size.
2. Dispersion—variation, spread, or scatter.
3. Skewness—asymmetry or lopsidedness.
4. Kurtosis—peakedness or relative influence of extreme deviations.

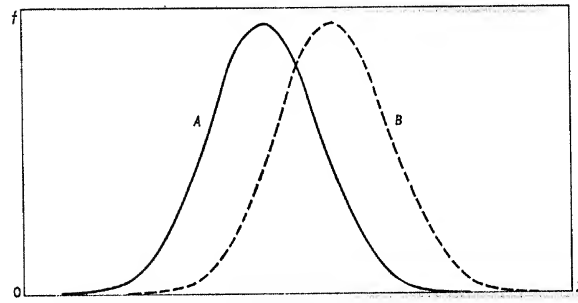
These four characteristics are illustrated in Chart 6-1 by smooth frequency curves. A frequency curve as defined in Chapter 4 portrays the frequency distribution of a population of continuous data in which the area under any segment of the curve corresponds to the number of values in that interval. Chart 6-1 is drawn so that the total area under each curve is unity and the area within any interval is equal to the relative frequency for that interval.

Suppose these curves represent the distribution of wage rates in a large factory. Panel 1 then shows that wages in department A *average* lower than those in department B, although both have the same dispersion. In panel 2, department A has a wider variation or *dispersion* of wages than department B, although both have the same average. The curves in both panels are symmetrical and normal. Panel 3 illustrates

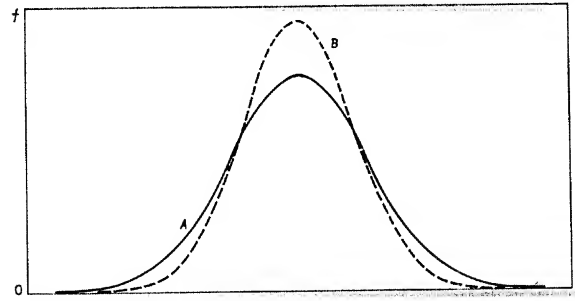
Chart 6-1

FOUR SUMMARY MEASURES OF A FREQUENCY DISTRIBUTION

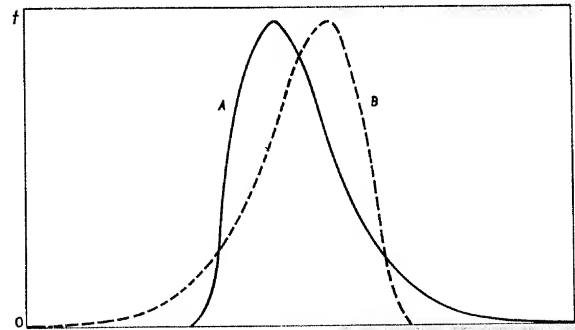
1. AVERAGE IS SMALL (A) OR LARGE (B)



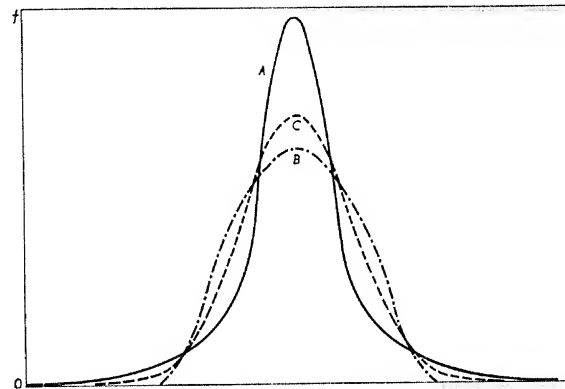
2. DISPERSION IS WIDE (A) OR NARROW (B)



3. SKEWNESS IS POSITIVE (A) OR NEGATIVE (B)



4. KURTOSIS IS PEAKED (A), FLAT-TOPPED (B) OR NORMAL (C)



skewness. Here most of the wages in department *A* are near the minimum rate, although some are much higher (i.e., skewness is positive or to the right); while in department *B* most of the wages are near the maximum (skewness is negative or to the left). Finally, panel 4 shows different types of *kurtosis* in three symmetrical distributions having the same average and the same dispersion (as measured by the standard deviation, to be explained later). The distribution in department *A* is peaked, since most of the workers receive about the same wage with few very high or low wages; while the distribution in department *B* is flat-topped, indicating that the typical wages cover a wider spread with fewer extreme deviations; and in department *C* the distribution is normal, as if it had been determined by chance.¹

Averages and measures of dispersion are the most important of these four kinds of summary measures. Dispersion will be described at length, and skewness very briefly, in this chapter. Kurtosis will be omitted, except for nontechnical references to the effects of extreme deviations.

PURPOSES OF MEASURING DISPERSION

Dispersion is the variation, or scatter, of a set of values. Measures of dispersion are needed for two basic purposes: (1) to gauge the reliability of averages and (2) to serve as a basis for control of the variability itself.

To illustrate the first purpose, suppose a company analyst is measuring the cost of living in a large city as one factor determining whether wages should be raised. If in five filling stations selected at random he finds that the price of standard gasoline varies between 33 and 34 cents per gallon, he might be justified in using the mean of as few as five prices, say 33.4 cents, to represent the price of gasoline. That is, the mean of five prices represents closely the price at each station, and it provides a reliable estimate of the mean price of all standard-grade gasoline sold in the city. On the other hand, prices of a certain type of woman's dress might vary from \$9.95 to \$24.95 in five stores. The mean of so few prices would then be highly unreliable as an estimate of the mean price of all such dresses in the city, but a measure of dispersion is needed to reveal this fact. To summarize the facts in most cases, therefore, *both an average and a measure of dispersion must be presented*.

When dispersion is small, the average is a typical value in that it closely represents the individual values, and it is reliable in that it is a good estimate of the corresponding average in the population. On the

¹ Curves *A*, *B*, and *C* are called leptokurtic, platykurtic, and mesokurtic, respectively.

other hand, when the dispersion is great, the average is not so typical and, unless the sample is very large, the average may be quite unreliable (see Chapter 11).

The second basic purpose of measuring dispersion is to determine the nature and causes of variation in order to control the variation itself. In matters of health, variations in body temperature, pulse beat, and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programs. Thus, measurement of dispersion is basic to the control of causes of variation.

Measures of dispersion include: (1) the range, (2) the quartile deviation, (3) the mean deviation, and (4) the standard deviation. These measures are analogous to the averages described in Chapter 5, both in their characteristics and methods of calculation.

THE RANGE

The range is the difference between the largest and the smallest values of a variable. It is the simplest of all measures of dispersion. For the gasoline prices varying from 33 to 34 cents per gallon, the range is 1 cent. The range can be easily computed in an array, but it cannot be determined accurately from a frequency distribution unless the high and low values in the end classes are known.

Sometimes the range is indicated merely by citing the largest and smallest figures themselves. Quotations of stock prices and commodity prices include the high and low for the day. Weather reports state the maximum and minimum temperatures. If the high and low values are not widely separated from the other values, as in these cases, the range may be a fairly good measure of dispersion. In particular, the range is the basic measure of variation used in quality control, as described in Chapter 25.

However, if the two extremes are erratic, the range is unreliable and misleading because it gives no hint of the dispersion of the intervening values. In the distribution of prices paid for cars, for example, the range might extend from a Rolls-Royce at \$20,000 to a used Jeep at \$800; this would give little information about the variation in prices paid by the majority of consumers. In general, if the population contains a few extreme deviations, the range obtained from a random sample is more unreliable than any other measure of dispersion. For these reasons, the range is not recommended for general use.

The influence of extreme deviations on a measure of dispersion can

be reduced by excluding a specified proportion of values at each end of the array and using the range of the remaining central values as the measure of dispersion. The simplest and most useful of these measures is the quartile deviation, which is explained below.

THE QUARTILE DEVIATION

The quartile deviation (Q) is defined by the formula

$$Q = \frac{(Q_3 - Q_1)}{2}$$

where Q_1 and Q_3 are the first and third quartiles, respectively. The quartiles are the three points which divide an array or frequency distribution into four roughly equal groups.² That is, the first or lower quartile, Q_1 , separates the lowest-valued quarter of the total number of values from the second quarter; the second quartile, Q_2 (almost always called the median), separates the second quarter from the third quarter; and the third or upper quartile, Q_3 , separates the third quarter from the top quarter. Consequently, the quartile range, $Q_3 - Q_1$, includes the middle half of the items. The quartile deviation is half this range.

The quartiles are widely used as measures of dispersion. *Dun's Review and Modern Industry*, for example, reports the medians and quartiles of 14 operating ratios in each of 71 types of manufacturers. Thus, the quartiles of net-profits-to-sales ratios for 56 drug manufacturers in 1965 were 2.97 and 9.57 percent, as compared with the median of 5.93 percent.³ This means that while the "typical" drug manufacturer earned 5.93 percent on sales, about one fourth of the companies earned less than 2.97 percent and one fourth earned over 9.57 percent, indicating a wide spread of profitability in this field. Similarly, the National Industrial Conference Board's *Management Record* reports the median and quartile salaries for various occupations by cities. In these cases, the quartiles themselves are reported rather than the quartile deviation.

Ungrouped Data

The first and third quartiles are found in an array just as is the median or second quartile. They are the values whose ranks or order

² The groups are rarely exactly equal, for reasons described under the median and because n is seldom a multiple of 4.

The term "quartile" is sometimes applied to an entire range of values rather than to a point. Thus, a score might be said to fall "in the upper quartile" (i.e., between the top value and the upper quartile partition point). Such a range, however, should be called "quarter" to avoid confusion with "quartile," which should refer only to a point.

³ *Dun's Review and Modern Industry*, November 1966, p. 74.

numbers are $n/4 + 1/2$ and $3n/4 + 1/2$, respectively, counting from the lowest value. Fractional order numbers are interpolated between neighboring values in the array.

In the case of the hourly earnings of 214 machine tool operators listed in Table 4-3, the value of Q_1 is the earnings which rank is $214/4 + 1/2$, or 54. This is the earnings of the 54th man,⁴ the middle man of the lower-paid half of the operators. Similarly, the value of Q_3 is the earnings of the man who is 161st from the bottom or 54th from the top, the middle man of the upper half. The values of Q_1 and Q_3 are found to be \$2.50 and \$2.70, respectively, from the original ungrouped data in Table 4-3. This means that about one fourth of the operators earn less than \$2.50, one fourth exceed \$2.70, and the middle half fall between these values. The quartile deviation is then $(2.70 - 2.50) \div 2$, or \$0.10.

Grouped Data

The quartiles can be estimated for a frequency distribution in the same way as the median by these analogous formulas:

$$Q_1 = L + \frac{i(n/4 - F)}{f} \quad Q_3 = L + \frac{i(3n/4 - F)}{f}$$

where L is the lower limit of the class containing the quartile; i is the class width; f is the frequency or number in that class; F is the cumulative frequency below that class; and n is the total number of values. In these formulas, it is assumed that values of X are spread evenly over each interval, as explained in connection with the median.

For the machine tool operators' earnings grouped in Table 6-1, Q_1 , the 54th value, falls in the third class ($L = \$2.45$, $f = 49$, $F = 25$); and Q_3 , the 161st value, falls in the fifth class ($L = \$2.65$, $f = 45$, $F = 137$). Therefore,

$$\begin{aligned} Q_1 &= 2.45 + .10(53.5 - 25) \div 49 \\ &= 2.45 + .10(.58) \\ &= 2.508 \text{ dollars per hour} \end{aligned}$$

$$\begin{aligned} Q_3 &= 2.65 + .10(160.5 - 137) \div 45 \\ &= 2.65 + .10(.52) \\ &= 2.702 \text{ dollars per hour} \end{aligned}$$

⁴ If there were 215 operators, Q_1 would rank $215/4 + 1/2$, or $54\frac{1}{4}$, i.e., one fourth of the way from the earnings of the 54th man to that of the 55th man from the bottom.

The quartile deviation is then $(2.702 - 2.508) \div 2 = .097$ dollars per hour. These three estimates check fairly closely with the exact values already obtained from the ungrouped data.

The quartiles can be located graphically from a cumulative frequency curve, or ogive (shown in Chart 4-4,) in the same manner as the median. To determine Q_1 , for example, draw a horizontal line from $n/4$

Table 6-1
INTERPOLATION FOR QUARTILES
IN A FREQUENCY DISTRIBUTION
HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL
OPERATORS

Lower Limit of Class (L)	Number in Class (f)	Number Earning Less (F)	Location of Quartiles
\$2.25	2	0	
2.35	23	2	
2.45	49	25	$Q_1 = \#54$
2.55	63	74	
2.65	45	137	$Q_3 = \#161$
2.75	25	182	
2.85	3	207	
2.95	4	210	
3.05	0	214	
Total	214		

on the Y axis to the "less than" curve; then drop a perpendicular and read off the value of Q_1 on the X axis.

The quartile deviation is relatively unaffected by extreme deviations. On the other hand, since the quartile deviation depends entirely upon the values of the quartiles Q_1 and Q_3 , its reliability depends on the degree of concentration at the quartiles of the population from which the sample is selected. In particular, if there are gaps in the population around the quartiles, the quartile deviation is unreliable. The measures of dispersion which follow differ from the quartile deviation in that they take into account the deviation of *every* value from the average.

THE MEAN DEVIATION

The mean deviation, sometimes called the average deviation, is exactly what its name implies. It is simply the mean of the absolute deviations of all the values from some central point, such as the arithmetic mean or median. The deviations must be averaged as if they were all positive, since the mean of plus and minus deviations would be zero

(if measured from the mean), or nearly so. The mean deviation theoretically should be measured from the median since it is then smallest, but it is usually more convenient to measure the deviations from the mean, as described below. There is little difference in the results.

The mean deviation is a concise and simple measure of variability. Unlike the range and quartile deviation, it takes every item into account, and it is simpler and less affected by extreme deviations than the standard deviation, which will be described in the next section. It is therefore often used in small samples that include extreme values. The National Bureau of Economic Research, for example, computes mean deviations to show how different business cycles vary in duration, intensity, and other respects: "The average deviations . . . bring out what we consider one of the most important aspects of cyclical behavior. Some economic processes are fairly uniform in their movements from cycle to cycle, and so have relatively small average deviations; most factors show wide diversity of movement, and so have large average deviations."⁵

Ungrouped Data

The formula for the mean deviation (measured from the arithmetic mean) in a set of ungrouped data is

$$MD = \frac{\sum |x|}{n}$$

where x is the *deviation* of each item X from the mean \bar{X} ; i.e., $x = X - \bar{X}$. The blinkers $| |$ mean that the signs are ignored. Then the sum (Σ) of the absolute deviations $|x|$ is divided by the number of values (n) to find the mean deviation (MD).

The mean deviation is computed in Table 6-2 for the price-earnings ratios of five electronics stocks, whose mean is 20.0. That is,

$$MD = \frac{\sum |x|}{n} = \frac{19.8}{5} = 4.0$$

This means that while the five price-earnings ratios averaged 20.0, there was a wide variation among them, since the average departure from the mean was 4.0. Furthermore, the sample includes only five stocks. Therefore, the average ratio of 20.0 must be considered rather unreliable as an estimate of the typical price-earnings ratio for electronics stocks generally, assuming a large population of such stocks.

⁵ Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946), p. 381.

Table 6-2
COMPUTATION OF MEAN DEVIATION
FOR UNGROUPED DATA
PRICE-EARNINGS RATIOS OF FIVE ELECTRONICS STOCKS

Common Stock	Price-Earnings Ratio (X)	Deviation from Mean $ x $
A.....	19.6	0.4
B.....	17.3	2.7
C.....	19.2	0.8
D.....	14.0	6.0
E.....	29.9	9.9
Total.....	100.0	19.8
Mean.....	20.0 = \bar{X}	4.0 = MD

Grouped Data

The mean deviation can be computed from grouped data by the formula

$$MD = \frac{\sum f|x|}{n}$$

where $|x|$ is the absolute deviation of the class midpoint (X) from the arithmetic mean, ignoring signs, and f is the frequency in that class.⁶ This formula will not be illustrated here, since its practical use is limited. The mean deviation has certain logical and mathematical limitations, such as disregarding plus and minus signs in averaging deviations. Consequently, the standard deviation is usually used instead for large distributions of grouped data.

THE STANDARD DEVIATION

The standard deviation is found by (1) *squaring* the deviations of individual values from the arithmetic mean, (2) summing the squares, (3) dividing the sum by $(n - 1)$, and (4) extracting the square root. Like the mean deviation, the standard deviation is based on the deviations of all values, but it is better adapted to further statistical analysis. This is partly because squaring the deviations makes them all positive, so that the standard deviation is easier to handle algebraically than the mean deviation. The standard deviation is therefore of such importance that it is, in fact, the "standard" measure of dispersion.

⁶ For a short-cut method of computing the mean deviation for grouped data, see Spurr, Kellogg, and Smith, *Business and Economic Statistics* (Homewood, Ill.: Richard D. Irwin, 1954), pp. 227-28.

Ungrouped Data

The basic formula for the standard deviation of ungrouped data is

$$s = \sqrt{\frac{\sum x^2}{n - 1}}$$

where s is the standard deviation; $x = X - \bar{X}$ is the deviation of any value of X from the arithmetic mean \bar{X} ; $\sum x^2$ is the sum of the squared deviations; and n is the number of items in the sample. The deviations may be squared most easily by referring to a table of squares, such as Appendix C or *Barlow's Tables*.

The square of the standard deviation is called the *variance*. This is an important concept in statistical inference, to be considered later.

The above formula is now commonly used in statistics because it provides the best estimate of the standard deviation of the population from which the sample was drawn. An alternative formula for the standard deviation is $\sqrt{\sum x^2/n}$, which measures the dispersion of the sample itself but tends to understate the dispersion of the population. Since we usually take a sample in order to estimate population values, we will use $n - 1$ in our equations for s , the sample standard deviation, and will regard s as an estimate of σ (small sigma), the population standard deviation. However, n may be substituted for $n - 1$ if desired; it makes little difference when n is large, as in most economic data.

For the five price-earnings ratios listed in Table 6-3, column 2, the

Table 6-3
COMPUTATION OF STANDARD DEVIATION
FOR UNGROUPED DATA
PRICE-EARNINGS RATIOS OF FIVE ELECTRONICS STOCKS

(1)	(2)	(3)	(4)	(5)
Common Stock	Price- Earnings Ratio (X)	Deviation from Mean ($x = X - \bar{X}$)	x^2	X^2
A.....	19.6	— .4	.16	384.16
B.....	17.3	— 2.7	7.29	299.29
C.....	19.2	— .8	.64	368.64
D.....	14.0	— 6.0	36.00	196.00
E.....	29.9	9.9	98.01	894.01
Total.....	100.0	.0	142.10	2,142.10
Mean.....	20.0			

deviations from the mean of 20.0 are shown in column 3 and the squares in column 4. Their sum (Σx^2) is 142.10, and $n = 5$ stocks. The standard deviation is then

$$s = \sqrt{\frac{\Sigma x^2}{n-1}} = \sqrt{\frac{142.10}{4}} = \sqrt{35.52} = 6.0$$

Short-Cut Method. While the above formula *describes* the standard deviation succinctly, it is usually easier to *compute* its value directly from the original data, without finding the deviations from the mean. The following formula can be used to give exactly the same result as the one above:

$$s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}}$$

In Table 6-3, column 5 shows the original X values squared for use in this formula; columns 3 and 4 are not needed. Then,

$$\begin{aligned} s &= \sqrt{\frac{2,142.10 - (100.0)^2/5}{4}} \\ &= \sqrt{\frac{2,142.10 - 2,000}{4}} \\ &= \sqrt{35.52} \\ &= 6.0 \end{aligned}$$

The standard deviation is larger than the mean deviation of 4.0. This is always true because the squaring of the deviations puts more emphasis upon the extreme items.

Grouped Data

In a frequency distribution the midpoint of each class is used to represent every value in that class. The basic formula for the standard deviation therefore becomes

$$s = \sqrt{\frac{\Sigma fx^2}{n-1}}$$

where x is the deviation of the class midpoint (X) from the arithmetic mean and f is the frequency in that class.

Short-Cut Method. The computation can be simplified by using the class midpoints (X) themselves rather than their deviations (x) from the mean, as follows:

$$s = \sqrt{\frac{\sum fX^2 - (\sum fX)^2/n}{n - 1}}$$

These two formulas are the same as those for ungrouped data except for using X as the class midpoint and f as the class frequency. A brief illustration is given in Table 6-4, which shows the prices of a transistor radio in six stores. The mean price is \$26.

Table 6-4
COMPUTATION OF STANDARD DEVIATION
FOR GROUPED DATA (TWO METHODS)
PRICES OF A TRANSISTOR RADIO IN SIX STORES

(1) Price in Dollars (Class Midpoint) X	(2) Number of Stores (Frequency) f	(3) Deviation from Mean (Dollars) x	(4) fx^2	(5) fX	(6) fX^2
27	2	1	2	54	1,458
26	3	0	0	78	2,028
25	0	-1	0	0	0
24	1	-2	4	24	576
Total	6		6	156	4,062

Using the first formula,

$$s = \sqrt{\frac{\sum fx^2}{n - 1}} = \sqrt{\frac{6}{5}} = 1.10 \text{ dollars}$$

Using the "short-cut" formula (which is not really shorter in this simple case),

$$\begin{aligned} s &= \sqrt{\frac{\sum fX^2 - (\sum fX)^2/n}{n - 1}} = \sqrt{\frac{4,062 - (156)^2/6}{5}} \\ &= \sqrt{\frac{4,062 - 4,056}{5}} = \sqrt{\frac{6}{5}} = 1.10 \text{ dollars} \end{aligned}$$

The results of the two formulas are thus identical. These methods are not discussed further because in practice the standard deviation of

grouped data is usually computed by a still shorter method, as described below.

A Short Method for Both Mean and Standard Deviation. The methods described above for computing the standard deviation and those discussed in Chapter 5 for computing the mean are quite arduous if the numbers are large. This section describes a shorter method for calculating both the mean and standard deviation for grouped data having class intervals of equal width.

This method is illustrated in Table 6-5. Although at first it may not

Table 6-5
COMPUTATION OF MEAN AND STANDARD DEVIATION
FOR GROUPED DATA—SHORTEST METHOD
HOURLY EARNINGS OF 214 APPRENTICE MACHINE TOOL OPERATORS

(1) Class Midpoint (Dollars) X	(2) Frequency f	(3) Deviation from Assumed Mean in Classes d	(4) fd	(5) fd^2
2.30	2	-3	-6	18
2.40	23	-2	-46	92
2.50	49	-1	-49	49
2.60	63	0	0	0
2.70	45	1	45	45
2.80	25	2	50	100
2.90	3	3	9	27
3.00	4	4	16	64
Total	214		19	395

appear shorter, a little practice will demonstrate that much time and labor can be saved because the multipliers are reduced to small whole numbers.

The steps for computing the mean and standard deviation by the short method are as follows.

1. List the class midpoints and the frequencies, as shown in columns 1 and 2 (Table 6-5).
2. Select any midpoint as the "assumed mean" (\bar{X}_a), preferably the midpoint of one of the middle intervals. In Table 6-5 the assumed mean is taken as \$2.60.
3. List the deviation (d) of each class midpoint from the assumed mean in units of the class interval, as in column 3. Thus, a zero is written opposite 2.60, the next larger midpoint is marked +1, the

next smaller -1 , and so on in whole numbers, 1, 2, 3, etc. Be sure to mark the deviations of the larger midpoints "+" and the smaller midpoints "-", irrespective of which end is listed first in the table. If there were a gap and then some values, say in the "3.15 and under 3.25" class, that class would have a deviation of 6, not 5, class units from the assumed mean.

4. Multiply the frequency in each class by its deviation and list the product (fd) in column 4, being sure to include the sign.
5. The total of column 4 is Σfd . Square this number to obtain $(\Sigma fd)^2$.
6. Multiply d (column 3) by fd (column 4) to obtain fd^2 (column 5). (Or square d and multiply by f .) Since the d 's are integers, column 5 can be easily calculated.

The formula for the arithmetic mean computed by the short method is

$$\bar{X} = \bar{X}_a + \frac{i\Sigma fd}{n}$$

where

\bar{X} = the arithmetic mean;

\bar{X}_a = the assumed mean placed at any class midpoint;

i = the width of the interval (measured from the lower limit of one interval to the lower limit of the next);

f = the frequency or number of items in each class;

d = the deviation of a midpoint from the assumed mean in class interval units;

Σfd = the sum of f times d for each class (not Σf times Σd); and

n = the total number of items.

In Table 6-5, therefore,

$$\begin{aligned}\bar{X} &= \bar{X}_a + \frac{i\Sigma fd}{n} \\ &= 2.60 + \frac{0.10(19)}{214} \\ &= 2.60 + 0.009 \\ &= 2.609, \text{ or } \$2.609 \text{ per hour}\end{aligned}$$

This method of computing the arithmetic mean yields precisely the same result as $\bar{X} = \Sigma fX/n$, the formula for the direct method.

Then compute the standard deviation from the formula

$$s = i \sqrt{\frac{\sum fd^2 - (\sum fd)^2/n}{n - 1}}$$

using the same symbols as above, and $\sum fd^2$ = the sum of f times d^2 for each class (the total of column 5).

Substituting the numbers from Table 6-5, the computation is

$$\begin{aligned} s &= i \sqrt{\frac{\sum fd^2 - (\sum fd)^2/n}{n - 1}} \\ &= 0.10 \sqrt{\frac{395 - (19)^2/214}{213}} \\ &= 0.10 \sqrt{\frac{395 - 1.69}{213}} \\ &= 0.10 \sqrt{1.85} \\ &= 0.136 \text{ dollars per hour} \end{aligned}$$

The result of this formula is the same as for the two other formulas for the standard deviation given, but the computations in columns 3, 4, and 5 are simpler. In any case, the mean and standard deviation for grouped data are slightly less exact than those computed from the original data, since in formulas containing f the values in each class are rounded off to the class midpoint.⁷

If the widths of class intervals in a frequency distribution are unequal, the class deviations must be adjusted to uniform units (such as the smallest interval or the highest common factor) in order to apply these short formulas. Otherwise the longer formulas should be used. If the distribution has an open end, neither the mean nor the standard deviation can be computed, unless the missing end values can be estimated.

⁷ The three formulas for grouped data would be exact if every value of X were equal to its class midpoint. In case the concentration of values tapers off on both sides of the mean, as in a normal distribution, it is appropriate to adjust for grouping errors by subtracting $i^2 \div 12$ from the variance. This is called *Sheppard's adjustment*. This adjustment is not generally recommended, however, because (1) when major points of concentration occur at midpoints the unadjusted formula is more nearly appropriate, (2) when values of X are evenly distributed over the intervals the one-twelfth adjustment should be *added*, not subtracted. Hence, the unadjusted formula is not only appropriate for one assumption but is also the mean of results obtained from two other assumptions. Finally, (3) errors of grouping are often small in comparison with other types of errors.

RELATION BETWEEN MEASURES OF DISPERSION

In a normal distribution there is a fixed relationship between the three most commonly used measures of dispersion. The quartile deviation is smallest, the mean deviation next, and the standard deviation σ is largest, in the following proportions:⁸

$$\begin{aligned} Q &\approx 2/3\sigma \\ MD &\approx 4/5\sigma \end{aligned}$$

where the sign \approx denotes approximate equality.

These relationships can be easily memorized because of the sequence 2, 3, 4, 5. The same proportions tend to hold true for many distributions that are not quite normal. They are useful in estimating one measure of dispersion when another is known or in checking roughly the accuracy of a calculated value. In the case of the machine tool operators, for example, $Q = \$0.097$, $MD = \$0.103$, and s , the estimate of σ , $= \$0.136$. Here σ could be estimated roughly from Q as $\sigma = 3/2Q = \$0.145$; or, more accurately from MD , as $\sigma = 5/4 MD = \$0.129$. If the computed standard deviation differs very widely from its value estimated from Q or MD , either an error has been made or the distribution differs considerably from normal.

Another comparison may be made of the proportion of items that are typically included within the interval of one Q , MD , or σ measured both above and below the population mean μ (small mu in Greek). In a normal distribution,

$$\begin{aligned} \mu \pm Q &\text{ includes 50 percent of the items} \\ \mu \pm MD &\text{ includes 57.51 percent of the items} \\ \mu \pm \sigma &\text{ includes 68.27 percent of the items} \end{aligned}$$

These relationships are shown graphically in Chart 6-2. Note that the standard deviation is the distance between the mean and the point of inflection on the normal curve, that is, the point where the curve changes from being concave downward to being concave upward, and where it is steepest.

For the machine tool operators, the interval around the sample mean $\bar{X} \pm Q$ is $\$2.609 \pm \0.097 , or from $\$2.512$ to $\$2.706$ per hour. This interval actually includes about 50 percent of the workers, and so the distribution is nearly normal in this respect. The proportions within the

⁸ More precisely, $Q = 0.6745\sigma$ and $MD = 0.7979\sigma$.

intervals $\bar{X} \pm MD$ and $\bar{X} \pm s$ are also nearly normal for the hourly earnings, since they contain 53 and 67 percent of the workers, respectively.

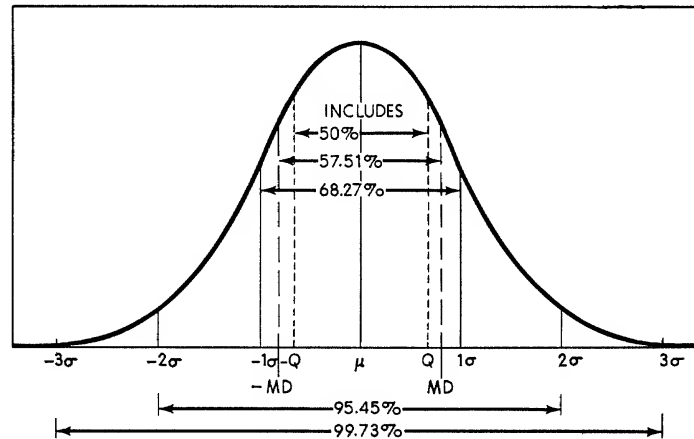
The proportions of items typically falling within 1, 2, and 3 standard deviations of the mean are also widely used in statistical analysis. In a normal distribution,

$$\begin{aligned}\mu \pm \sigma & \text{ includes 68.27 percent of the items} \\ \mu \pm 2\sigma & \text{ includes 95.45 percent of the items} \\ \mu \pm 3\sigma & \text{ includes 99.73 percent of the items}\end{aligned}$$

These relations are also shown graphically in Chart 6-2. The interval $\bar{X} \pm 2\sigma$ thus includes about 19 out of 20 of the items, while $\bar{X} \pm 3\sigma$ includes nearly all of them. In the case of the machine tool operators,

Chart 6-2

PROPORTIONS OF AREA OF NORMAL CURVE INCLUDED IN INTERVALS
BASED ON COMMON MEASURES OF DISPERSION



the interval $\$2.609 \pm (3 \times \$0.136)$, or from $\$2.201$ to $\$3.017$, includes 212 out of 214 workers (Table 4-3). In general, so long as the departure from symmetry is only moderate, an interval of 3σ on both sides of the average will give the practical limits of the distribution.

Which Measure of Dispersion to Use?

As in the case of averages, the selection of the proper measure of dispersion depends on three main factors:

1. The concept of dispersion required by the problem. Is a single pair of values adequate, such as the two extremes or the two quartiles

(range or Q)? Or is a simple average of all absolute deviations from the mean or median needed (i.e., mean deviation)? Or an average (the standard deviation) that is better adapted for further calculations?

2. The type of data available. If they are few in number, or contain extreme values, avoid the standard deviation. If they are generally skewed, avoid the mean deviation as well. If they have gaps around the quartiles, the quartile deviation should be avoided.
3. The peculiarities of the dispersion measures themselves. These are summarized under "Characteristics of Measures of Dispersion," below.

As a rule of thumb, the median and quartiles may be used as simple, easily understandable summary values for rough or skewed data, as in a distribution of personal incomes, but the overall range should be avoided.⁹ The mean deviation is commonly used to give equal weight to all deviations where n is small and in ungrouped data, even if the distribution is somewhat erratic, as in time series. But if n is large and the distribution is fairly symmetrical, and if more refined analysis is needed, such as the study of inference or correlation, the standard deviation should be used instead. A major reason for the widespread use of the standard deviation is that it has the smallest sampling error of any dispersion measure when the distribution is normal; that is, the sample value tends to deviate from the population value by the smallest percentage.

Characteristics of Measures of Dispersion

The characteristics of the individual measures of dispersion are summarized below:

Range:

1. The range is the easiest measure to compute and to understand, but
2. It is often unreliable, being based on two extreme values only.

Quartile Deviation:

1. The quartile deviation is also easy to calculate and to understand.
2. It depends on only two values, which include the middle half of the items.
3. It is usually superior to the range as a rough measure of dispersion.

⁹ An exception is the use of the range in quality control, discussed in Chapter 25.

4. It may be determined in an open-end distribution, or one in which the data may be ranked but not measured quantitatively.
5. It is also useful in badly skewed distributions or those in which other measures of dispersion would be warped by extreme values.
6. However, it is unreliable if there are gaps in the data around the quartiles.

Mean Deviation:

1. The mean deviation has the advantage of giving equal weight to the deviation of every value from the mean or median.
2. Therefore, it is a more sensitive measure of dispersion than those described above and ordinarily has a smaller sampling error.
3. It is also easier to compute and to understand and is less affected by extreme values than the standard deviation.
4. Unfortunately, it is difficult to handle algebraically, since minus signs must be ignored in its computation.

Standard Deviation:

1. The standard deviation is usually more useful and better adapted to further analysis than the mean deviation.
2. It is more reliable as an estimator of the population value than any other dispersion measure, provided the distribution is normal.
3. It is the most widely used measure of dispersion and the easiest to handle algebraically.
4. However, it is harder to compute and more difficult to understand, and
5. It is greatly affected by extreme values that may be due to skewness of data.

MEASURES OF RELATIVE DISPERSION

The measures of dispersion so far described are expressed in original units, such as dollars. These values may be used to compare the variation in two distributions provided the variables are expressed in the same units and are of about the same average size. In case the two sets of data are expressed in different units, however, such as tons of coal versus cubic feet of gas, or if the average size is very different, such as executives' salaries versus laborers' wages, the absolute measures of dispersion are not comparable and measures of *relative* dispersion should be used instead.

A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average and is usually expressed as a

percent. It is sometimes called a *coefficient of dispersion* because "coefficient" means a ratio or pure number that is independent of the unit of measurement. A coefficient of dispersion may be computed from either the quartile or mean deviation¹⁰ but is usually expressed as the ratio of the standard deviation to the mean, s/\bar{X} .

Thus, for the apprentice machine tool operators' earnings, the coefficient of dispersion is

$$s/\bar{X} = 0.136/2.609 = 5.2 \text{ percent}$$

That is, the standard deviation is 5.2 percent of the mean earnings. If a group of foremen had a standard deviation of \$.160 and mean earnings of \$8.00 an hour, their earnings would vary more than those of the operators in dollars, to be sure (\$.160 versus \$.136), but they would vary less relative to their average earnings ($0.160 \div 8.00 = 2.0$ percent versus 5.2 percent). The *relative* measure is the more significant comparison.

Standard Deviation Units

Individual deviations from the mean ($x = X - \bar{X}$) may also be reduced to comparable units by dividing them by the standard deviation (s). Thus, for a machine tool operator earning \$2.80 an hour, or \$0.191 above the mean of \$2.609, $x/s = 0.191/0.136 = 1.40$. His wage is, therefore, 1.40 standard deviations above the mean, a value which is comparable with, say, his output in units produced, which may be 2.20 standard deviations above the mean. Perhaps he rates a raise in pay!

The values of x/s will vary from approximately +3 to -3 for any set of data, since this spread includes nearly all the items in a normal distribution. The interval $\bar{X} \pm 3s$ therefore provides the practical limits of variation used in quality control and other applications. Variation greater than these limits indicates the presence of abnormal forces that must be isolated and eliminated.

SKEWNESS

Skewness means the lack of symmetry in the shape of a frequency curve. The extent of this lopsidedness is another important characteristic of a frequency distribution.

The simplest measure of skewness is based on the spread between the arithmetic mean and median. They are identical in a symmetrical distri-

¹⁰ The formulas are $(Q_3 - Q_1)/(Q_3 + Q_1)$ and MD/\bar{X} , respectively.

bution. In a skewed distribution, however, the mean is pulled out in the direction of the extreme values while the mode remains under the highest point of the curve, and the median, which is affected by the number of extreme values but not their value, tends to fall about one third of the way from the mean toward the mode, provided the skewness is moderate.

A *coefficient of skewness* may therefore be defined as follows:

$$Sk = \frac{3(\bar{X} - Md)}{s}$$

where \bar{X} is the mean; Md is the median; and s is the standard deviation.

The numerator $3(\bar{X} - Md)$ is used instead of $(\bar{X} - \text{Mode})$ because the mode is often difficult to locate accurately. Dividing by s expresses the measure in standard deviation units, so that it is comparable between distributions that differ in unit of measurement or in average size. If the mean exceeds the median, the skewness is positive; otherwise it is negative.

The formula will not be illustrated here because of its limited practical use. The accurate measurement of skewness requires more advanced techniques. In elementary analysis, skewness is ordinarily treated in descriptive terms rather than being summarized by a single measure.

USES OF MEASURES OF DISPERSION

As the student gains experience with the analysis of data, he will perceive opportunities for the use of measures of dispersion other than those which have just been described. The following summary briefly indicates these various applications.

Aid in Description

The simplest and most common use of a measure of dispersion is in the description of data. Averages are typical values, but measures of dispersion indicate the scatter of the data. The extent and direction of skewness should also be noted.

Comparison of Dispersion

The average values of two sets of data may be very similar, while the range and pattern of scatter differ greatly. If the data are generally alike, the measures of dispersion can be compared in absolute units to determine how the data differ in their variability. When several sets of data are expressed in different kinds of units or in similar units of widely

different size, comparisons based on measures of relative dispersion are usually more appropriate.

Provision of a Standard

By the use of measures of dispersion, particularly the standard deviation, it is possible to compare the variation in a given group of data with that of the normal curve as a standard. It has been pointed out that approximately 68 percent of all the items in a normal distribution are included between one standard deviation above the mean and one standard deviation below the mean. When characteristics of a variable are expressed in standard deviation units, its distribution can be compared with a normal distribution. This use is at the very heart of studies of reliability of sample averages, quality control programs in industrial production, and other applications of statistical methods.

Measurement of Sampling Errors

Reliability of sample averages is an important part of statistical analysis. Averages vary by chance from sample to sample in the same population. In order to evaluate the reliability of the average in a single sample, we must know more about the variation of that average in all possible samples. The standard deviation is ordinarily used in this type of study, as explained in Chapter 11.

SUMMARY OF FORMULAS

Since the characteristics of the various measures of dispersion and skewness have been summarized above, the chapter may be concluded by listing the principal formulas used:

<i>Measure</i>	<i>Ungrouped Data</i>	<i>Grouped Data</i>
Range.....	Subtract end values	Same
Quartile deviation.....	$Q = \frac{Q_3 - Q_1}{2}$	Same
	Q_1 is $\#n/4 + 1/2^*$	$Q_1 = L + \frac{i(n/4 - F)}{f}$
	Q_3 is $\#3n/4 + 1/2^*$	$Q_3 = L + \frac{i(3n/4 - F)}{f}$
Mean deviation.....	$MD = \frac{\sum x }{n}$	$MD = \frac{\sum f x }{n}$
Standard deviation.....	$s = \sqrt{\frac{\sum x^2}{n - 1}}$	$s = \sqrt{\frac{\sum fx^2}{n - 1}}$

* In an array, counting from lowest value.

$$\text{Short-cut method} \dots s = \sqrt{\frac{\sum X^2 - (\sum X)^2/n}{n-1}} \quad s = \sqrt{\frac{\sum fX^2 - (\sum fX)^2/n}{n-1}}$$

Shorter method for mean and standard deviation, if data are grouped into classes of equal width:

$$\text{Mean} \dots \bar{X} = \bar{X}_a + \frac{i \sum fd}{n}$$

$$\text{Standard deviation} \dots s = i \sqrt{\frac{\sum fd^2 - (\sum fd)^2/n}{n-1}}$$

Relative dispersion . . . Divide measure of absolute dispersion by appropriate average, e.g., s/\bar{X} .

$$\text{Skewness} \dots Sk = \frac{3(\bar{X} - Md)}{s}$$

PROBLEMS

1. Cite actual or hypothetical illustrations, not given in the text, of each of the following:
 - a) Two main purposes of measuring dispersion.
 - b) Positive and negative skewness.
 - c) Narrow dispersion and peaked kurtosis.
2. The values below show the number of hours of operation before repairs were required for eight power lawn mowers:

No. of Hours

21
27
29
35
29
21
27
35

Total = 224 hours

Compute and explain briefly the meaning of:

- a) The third quartile.
 - b) The mean deviation.
 - c) The standard deviation.
 - d) A measure of relative dispersion, using the standard deviation.
 - e) The largest value (35) expressed in standard-deviation units above the mean.
 3. In Chapter 4, Problem 12:
 - a) Find the range and quartile deviation from your original list of 112 items.
 - b) Interpolate the quartiles and compute the quartile deviation from your frequency distribution of these data.
 - c) Why do the quartile values differ in (a) and (b)?
 4. Using your frequency distribution in the problem above:
-

- a) Compute the standard deviation.
 - b) Explain the meaning of this measure in terms of electronic workers' earnings.
 - c) Should this value of s differ from the following? Give reasons.
 - (1) The s of the original ungrouped data.
 - (2) The s for the other formulas containing f .
 - d) Estimate the mean deviation from the standard deviation, assuming a nearly normal distribution.
5. Answer the same questions as in Problem 4, above, for the starting salaries in whichever of the five fields is assigned in Chapter 4, Problem 9.
6. A purchasing agent obtained samples of incandescent lamps from two suppliers. He had the samples tested in his own laboratory for length of life, with the following results:

LENGTH OF LIFE IN HOURS	SAMPLES FROM	
	Company A	Company B
700 and under 900.....	10	3
900 and under 1,100.....	16	42
1,100 and under 1,300.....	26	12
1,300 and under 1,500.....	8	3
Total.....	60	60

- a) Which company's lamps have the greater average length of life?
 - b) Which company's lamps are more uniform?
7. a) What ratio is MD to Q , in a normal distribution?
- b) The interval $\mu \pm 3\sigma$ includes nearly all the items in a normal distribution. Express this range in Q units.
- c) If you compute the standard deviation to be 0.612 pounds, and note as a rough check that the overall range is 36 pounds, what is the most obvious type of error you might have made?
- d) In a normal distribution of test scores with $\mu = 60$, $\sigma = 9$, what percent of scores exceeds 33? 51? 78?
8. If a test of 100 pieces of cotton thread shows a mean breaking strength of 15 pounds and a median breaking strength of 14.8 pounds, with a standard deviation of 3 pounds, about what number of pieces of thread in the lot should have a breaking strength between 12 and 21 pounds?
9. Regarding the dimensions of 63 gears in Table 4-2.
- a) Estimate the standard deviation of the whole lot from which this sample was drawn.
 - b) Check your result against the rough estimate of σ as one sixth the range (since the interval $\bar{X} \pm 3\sigma$ includes practically all items in a normal distribution).
 - c) How far does the largest gear (0.4270) differ from the mean in standard deviation units?

10. In Chapter 4, Problem 13:
- Compute whatever measure of dispersion you think most appropriate and explain its significance.
 - If there are any dispersion measures you cannot compute from these data, name them and indicate why you cannot.
11. In Chapter 4, Problem 14:
- Compute the standard deviation.
 - Find the estimated variance for all such cars. Explain its significance.
 - If you get 14 miles per gallon with this car, how many standard deviations are you below the mean of 18.82 miles per gallon?
12. In Chapter 5, Problem 13:
- Estimate the quartile deviation of refrigerator ages to the nearest year.
 - Is the distribution of refrigerator ages normal, negatively skewed, open-ended, or bimodal?
13. A firm which services household appliances for a national manufacturer is trying to determine where it should locate a service facility and its fleet of service trucks. The territory to be serviced lies along a straight highway and includes nine cities of roughly equal size. (See the sketch.) The manager decided to use the mean distance (counting the north end of the territory as zero as the location for the facility and the truck fleet. Thus, he has decided upon City F for the facility (mean = $225/9 = 25$).

MAP OF SERVICE TERRITORY		
Miles from City A		
0	•	City A
5	•	City B
10	•	City C
15	•	City D
20	•	City E
25	•	City F
40	•	City G
50	•	City H
60	•	City I
Total		225

- Compute the mean deviation of miles from the mean.
 - What does this figure tell the manager about the distance his service trucks will have to travel?
 - Before the manager has found a location, an assistant suggests that perhaps the median is a better measure to use here. Accordingly, the assistant suggests that City E, which is the middle city at 20 miles on the scale, be chosen as the site. Compute the mean deviation *about the median* (20).
 - By comparing this with the answer to (a) above, determine in which city the facility should be located. Why?
 - Do you think there is any better location? Explain.
14. As a further step in your analysis you wish to compare the dispersion of burning life for the two brands of electron tubes described in Chapter 4, Problem 15. The following calculations have been made from the raw data:

	BRAND A	BRAND B
ΣX	25,525	17,825
ΣX^2	6,888,125	4,999,375
n	120	80
\bar{X}	212.71	222.81

- Calculate the standard deviation for each brand of tube.
- Estimate the quartile deviation for each distribution from your cumulative frequency curve [Chapter 4, Problem 15 (d)].
- Compare the dispersion of the two distributions using both measures. Which measure gives the best general description in this case? Why?
- In Chapter 4, Problem 15 (d) you estimated the medians graphically. Using this estimate and the means above, what can you say about the skewness of these distributions?

SELECTED READINGS

CROXTON, FREDERICK E., AND COWDEN, DUDLEY J. *Practical Business Statistics*. 3d ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1960.

Chapters 8 and 15 to 18 provide a detailed treatment of ratios, frequency distributions, averages, and dispersion.

FREUND, JOHN E., AND WILLIAMS, FRANK J. *Modern Business Statistics*, chaps. 2 to 5. Englewood Cliffs, New Jersey: Prentice-Hall, 1958.

A clear discussion of frequency distributions, measures of "location," and of variation (more detailed than in their *Elementary Business Statistics*, 1964).

LEABO, DICK A. AND SMITH, C. FRANK. *Basic Statistics for Business and Economics*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1964.

Chapters 2 to 4 contain a careful account of basic statistical measures.

MEYERS, CECIL H. *Elementary Business and Economic Statistics*, chaps. 2 to 4. Belmont, California: Wadsworth Publishing, 1966.

A detailed discussion of frequency distributions, averages and dispersion.

NEISWANGER, WILLIAM A. *Elementary Statistical Methods*, chaps. 8 to 10. Rev. ed. New York: Macmillan, 1956.

A careful analysis of frequency distributions, averages, dispersion, and skewness.

NETER, JOHN, AND WASSERMAN, WILLIAM. *Fundamental Statistics for Business and Economics*, chaps. 3, 5, and 6. 3d ed. Boston: Allyn & Bacon, 1966.

Includes analysis of relationships by cross classification of data, as well as ratios and frequency distribution analysis.

U.S. GOVERNMENT, SMALL BUSINESS ADMINISTRATION. *Ratio Analysis for Small Business*. Washington, D.C.: Superintendent of Documents, 1957.

A popular survey of principal business ratios, sources of published ratios, and their analysis and evaluation.

YULE, G. UDNY, AND KENDALL, M. G. *An Introduction to the Theory of Statistics*. 14th ed. London: Charles Griffin, 1950.

Chapters 5 to 7 provide a comprehensive treatment of frequency distributions, averages, dispersion, skewness, and kurtosis.

7. AN INTRODUCTION TO PROBABILITY THEORY

PROBABILITY THEORY is a branch of mathematics that is eminently useful to the businessman. To a great extent, statistics is built upon the foundations of probability. The evaluation of information obtained from samples depends upon probability theory for its interpretation. Also, the businessman—like the poker player or military strategist—must make decisions in the face of uncertainty as to the future. He can express his judgment by attaching a numerical probability to each possible event that might affect the outcome of his decisions, and he can use these probabilities, together with economic information, to improve his decision-making process.

BASIC CONCEPTS

A *probability* is a number between 0 and 1, inclusive, representing the chance or likelihood that an event will occur. A probability of zero ($P = 0$) means the event is impossible; if $P = 0.50$, there is "half a chance" that it will occur; if $P = 1$, the event is certain to occur. The value of P cannot be negative or greater than one.

A probability may be thought of as the relative frequency of "successes" (i.e., the occurrence of a certain event) in a random process over a great number of trials. Relative frequency is the number of successes divided by the number of trials. Suppose we roll dice, and define a success as throwing an ace (1). If the dice are "fair," the six faces 1 through 6 are equally likely, and the ratio of aces to total throws will approach $1/6$ in the long run. We then define the probability of throwing an ace as $1/6$. The process of shooting dice is a random one because we do not know in advance the outcome of any given roll. In

general, if r is the number of successes in n trials, then the limit of r/n for larger and larger values of n is defined as the *probability of success in a single trial*.

Sources of Probabilities

The theoretical concept given above is difficult to apply in practice, but we can *estimate* probabilities in any of three ways:

1. **Relative Frequency of Past Events.** Probabilities can be estimated from relative frequencies either in a controlled experiment or in a sample survey of a large, finite population. To illustrate an experiment, suppose we set up a machine to turn out a new part and conduct an extended test run in which 5 percent of the parts prove to be defective. Then, if the process is controlled so that there is no change in quality of output, we can say that the probability is 0.05 that the next part will be defective. Of course, this part will in fact be either defective or good; our prior probability is derived from the long-run experience with many parts.

The probabilities for more complicated events can be determined from the probabilities for much simpler events by means of *simulation*—using an experimental model designed to approximate actual conditions. In studying an inventory system, for example, the orders of customers, the stock available, and the time necessary to replenish stocks would be incorporated in the model. A customer order is initiated and its effect is traced upon the inventory system. This is repeated for other orders and the behavior of the inventory system determined (e.g., the probabilities that demand will exceed supply by 0, 1, 2, . . . items, respectively). Simulation is described in Chapter 17.

Probabilities can also be estimated from the relative frequency with which an event occurs in a sample survey of a large finite population. Thus, in Table 4-4, the survey of machine tool operators reveals that 29 percent of the total earn about \$2.60 an hour. Then, the estimated *probability* is 0.29 that an operator drawn at random from the whole group of such operators would earn about \$2.60. Similarly, the probabilities for men and women buyers in the next section are based on their relative frequency in the sample survey cited.

2. **Theoretical Distributions.** In some situations, probabilities can be determined without recourse to relative frequencies. Thus, in rolling dice, we can state the probability of an ace as $1/6$ without actually rolling a die, simply because the six faces are equally likely to turn up. The probabilities for complicated events, too, can be derived from simple assumptions. For example, in tossing a fair coin four times, the

probabilities of from 0 to 4 heads may be derived from the fact that the probability of a head on one flip is $1/2$. The probability is $1/16$ for no heads, $1/4$ for one head, etc., as listed in Table 7-8. Such probabilities can be determined from the binomial distribution described in Chapter 8, without recourse to experiments or surveys based on past experience. The validity of such theoretical distributions depends upon how closely the assumptions match the real-world situation. (For example, the probabilities in Table 7-8 do not apply if, in fact, our coin is bent.)

3. Subjective Judgment. If none of these methods can be used, the decision-maker must estimate probabilities on the basis of his judgment and experience. An automobile manufacturer may judge the chances to be two out of three that customers will prefer one body style over another. The weatherman may say: "The chances are 6 out of 10 for rain." Most betting odds on athletic events are set by personal judgment. To include these situations, we enlarge the definition to include *subjective probability*. A subjective probability is an evaluation by a decision-maker of the relative "likelihood" of unknown events.¹ It is his betting odds on the occurrence of the event. Since it is personal to the decision-maker, two individuals may attach different subjective probabilities to the same event. Even so, these subjective probabilities can be used in decision-making in the same manner as the more objective probabilities described above.

Joint, Marginal, and Conditional Probabilities

Before proceeding, it is necessary to establish certain definitions. This can be done best by illustration. In studying the buying behavior of customers of a certain product, suppose you have taken the following random sample of 1,000 customers entering a department store:

Table 7-1
BUYING BEHAVIOR OF 1,000 MEN AND WOMEN
(Percent of Total)

	Men (M)	Women ($\sim M$)	Total
Buyer (B).....	3	17	20
Nonbuyer ($\sim B$).....	27	53	80
Total.....	30	70	100

¹We could be more precise and define subjective probability in terms of decision-makers' preferences for hypothetical lotteries. For our purposes, the intuitive definition above will suffice. For more detail, see Chapters 1 to 5 in the Pratt, Raiffa, and Schlaifer reference listed at the end of Chapter 8.

Suppose we are going to pick a customer from this group by chance. Then:

1. **Simple Probability.** Probability of drawing a man: $P(M) = 0.30$. The symbol $P(A)$ is used to denote the probability of an event A . The event "not- A " is represented by $\sim A$. Thus, the simple probability of drawing a woman is $P(\sim M) = 0.70$.

2. **Joint Probability.** The probability of getting a customer with two (or more) specific characteristics. For example, the probability of drawing a customer who is both a buyer and a man is $P(B, M) = 0.03$, and the probability of drawing a customer who is a woman, nonbuyer is $P(\sim M, \sim B) = 0.53$.

3. **Marginal Probability (on the margin of the table).** The total probability of drawing a man, made up of the probability of men buyers plus the probability of men nonbuyers.

$$P(M) = P(M, B) + P(M, \sim B) = 0.03 + 0.27 = 0.30$$

Marginal probability is no more than simple probability viewed in a different light. That is, simple probability is a singular concept, whereas the marginal probability is essentially a sum of joint probabilities.

4. **Conditional Probability.** Suppose that we know that the customer drawn was a man. Given this information, what is the probability that he is also a buyer? This is the conditional probability $P(B | M)$. The symbol $P(B | M)$ is read as the probability of a buyer *given* a man. Since 30 percent of the customers are men and 3 percent are buyers, $P(B | M) = 0.03/0.30 = 0.10$. From the above illustration, we can determine the general rule or mathematical definition of conditional probability:

Conditional probability of B given M :

$$P(B | M) = \frac{P(B, M)}{P(M)} = \frac{\text{joint probability of } B \text{ and } M}{\text{marginal probability of } M}$$

From this definition we can find, for example, the probability of a buyer, given that the customer is a woman:

$$P(B | \sim M) = \frac{P(B, \sim M)}{P(\sim M)} = \frac{0.17}{0.70} = 0.243$$

On the other hand, consider $P(M | B)$, the probability of the customer being a man, given that he is a buyer:

$$P(M | B) = \frac{P(B, M)}{P(B)} = \frac{0.03}{0.20} = 0.15$$

Note that this is not equal to $P(B | M)$ above.

As another illustration, suppose that we had an ordinary deck of cards. The cards can be classified as follows:

Table 7-2

PROBABILITIES IN DRAWING CARDS			
	Red Card, R	Black (nonred), $\sim R$	Total
Honor (A, K, Q, J, 10).. H	10/52	10/52	20/52 = 10/26
Nonhonor..... $\sim H$	16/52	16/52	32/52 = 16/26
Total.....	26/52 = 1/2	26/52 = 1/2	1

Now:

Simple Probability. The probability of drawing a red card, $P(R) = 1/2$.

Joint Probability. The probability of drawing a black honor, $P(H, \sim R) = 10/52$.

Marginal Probability. The probability of drawing a red card, viewed as the sum of the probabilities of red honors and red nonhonors,

$$P(R) = P(H, R) + P(\sim H, R) = 10/52 + 16/52 = 1/2.$$

Conditional Probability. The probability of an honor, given that we have drawn a red card,

$$P(H | R) = \frac{P(H, R)}{P(R)} = \frac{10/52}{26/52} = 10/26$$

Note that the simple probability of drawing an honor is also the same, that is, $P(H) = 10/52$. Hence, our knowledge that the card was red gave us no additional information about whether or not it was an honor, since the probabilities were exactly the same. This property is known as *statistical independence*.

Definition of Statistical Independence

When $P(H | R) = P(H)$, we say that the events H and R are *statistically independent*. That is, the event H is just as likely to occur when event R occurs as it is when event $\sim R$ occurs. (There is the same fraction of red honors as black honors.) Statistical independence implies

that knowledge of one event is of no value in predicting the occurrence of the other event.

To illustrate the notion of statistical independence, let us carry out the example of the buying behavior of customers and classify customers by age as well as sex. We could have the following table:

Table 7-3
BUYING BEHAVIOR OF 1,000 MEN AND WOMEN, BY AGE
(Percent of Total)

	Men (M)		Women ($\sim M$)		Total
	Young (Y)	Older ($\sim Y$)	Young (Y)	Older ($\sim Y$)	
Buyer (B).....	1	2	4	13	20
Nonbuyer ($\sim B$).....	5	22	15	38	80
Total.....	6	24	19	51	100

And the reader can easily verify that

$$\begin{array}{ll} \text{Total men} &= 30\% \\ \text{Total women} &= 70\% \end{array} \quad \begin{array}{ll} \text{Total young} &= 25\% \\ \text{Total older} &= 75\% \end{array}$$

Now, the simple probability of a buyer is $P(B) = 0.20$. The marginal probability of a young person is

$$\begin{aligned} P(Y) &= P(B, M, Y) + P(\sim B, M, Y) + P(B, \sim M, Y) \\ &\quad + P(\sim B, \sim M, Y) \\ &= 0.01 + 0.05 + 0.04 + 0.15 = 0.25 \end{aligned}$$

The conditional probability of a buyer given a young person is

$$P(B | Y) = \frac{P(B, Y)}{P(Y)} = \frac{0.01 + 0.04}{0.25} = 0.20$$

Note that this conditional probability equals the simple probability of a buyer, $P(B)$. Hence, age and buying behavior are independent. Knowledge of age is of no value in predicting whether or not a person is a buyer. The fact that age and buying behavior are independent also implies that

$$\begin{aligned} P(\sim B | Y) &= P(\sim B); \quad P(B | \sim Y) = P(B); \\ &\text{and } P(\sim B | \sim Y) = P(\sim B) \end{aligned}$$

Buying behavior and sex are not independent, however. Recall that the probability of buyer, given man, is $P(B|M) = 0.10$. But the probability of a buyer is $P(B) = 0.20$. Hence, B and M are not independent. Knowledge of the sex of a customer gives us a better probability estimate as to whether the person will be a buyer. (Men are less likely to buy than women.)

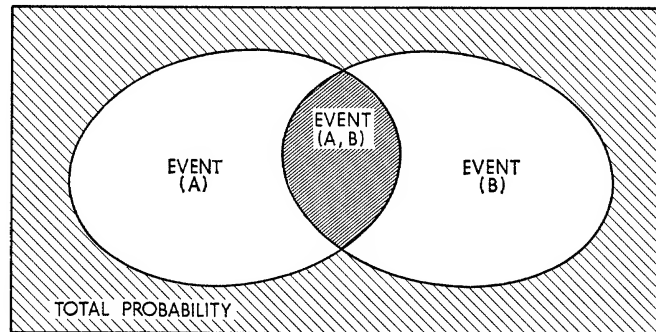
RULES FOR DEALING WITH PROBABILITIES

Addition of Probabilities

A set of events are said to be *mutually exclusive* if the occurrence of one excludes the occurrence of any of the others. For example, in

Chart 7-1

PROBABILITY OF NONMUTUALLY EXCLUSIVE EVENTS



drawing cards from a deck, the occurrence of the event “draw of a king” eliminates the possibility of the event “draw of a queen.” Hence, the events are mutually exclusive.

If a set of events are mutually exclusive, the probability of one or another of the events occurring is the sum of probabilities of the events occurring individually. Thus, if events A and B are mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B)$$

This is known as the *addition rule* for probabilities. Actually, the rule is fairly obvious; we have used it several times without stating it. For example, the probability of drawing a spade from a deck of cards is $1/4$. The probability of drawing a spade or a heart is $1/4$ plus $1/4$ or $1/2$.

If two events A and B are not mutually exclusive, then there is some probability that both can occur. The area of overlap is precisely the joint probability $P(A, B)$, as illustrated in Chart 7-1. This area is counted

twice in the addition formula used above for mutually exclusive events. We can modify the formula to obtain the *addition rule for events that are not mutually exclusive*:

$$P(A \text{ or } B) = P(A) + P(B) - P(A, B)$$

In the example illustrated in Table 7-1, the events "buyer" and "man" are *not* mutually exclusive since there are male buyers; that is, the event "buyer" does not rule out the possibility of the event "man." Hence, the probability of a man or buyer is

$$\begin{aligned} P(M \text{ or } B) &= P(M) + P(B) - P(M, B) \\ &= 0.30 + 0.20 - 0.03 = 0.47 \end{aligned}$$

A set of events is said to be *collectively exhaustive* if all possible occurrences are included. For example, the set of events "drawing a red card" and "drawing a black card" are collectively exhaustive; there are no other possibilities. The set of events "man," "buyer," and "woman nonbuyer" are collectively exhaustive (though not mutually exclusive).

The sum of the probabilities for a set of mutually exclusive and collectively exhaustive events equals one. This follows from the addition rule and from the fact that some event must occur.

Multiplication of Probabilities

The rule for multiplication of probabilities is merely an extension of the definition of conditional probability. The joint probability that *both* events A and B will occur equals the conditional probability of A given B , times the probability of B . In symbols,

$$P(A, B) = P(A | B) P(B)$$

As examples, consider the following:

If we knew that the probability of a man customer is $P(M) = 0.30$, and the probability that a man customer will be a buyer is $P(B | M) = 0.10$, the probability that a customer will be both a buyer and a man is

$$P(B, M) = P(B | M) P(M) = 0.30 \times 0.10 = 0.03$$

Suppose there were three balls in an urn, two white and one black. What is the probability of drawing both of the white balls in two draws (without putting the first ball back)?

Probability of white on first draw = $P(W_1) = 2/3$

Probability of second white, given first white = $P(W_2 | W_1) = 1/2$

Hence, the probability of a first white and a second white is

$$P(W_1, W_2) = P(W_2 | W_1) P(W_1) = 1/2 \times 2/3 = 1/3$$

Multiplication of Probabilities for Independent Events. When events are independent, $P(A | B) = P(A)$ and hence the rule becomes $P(A, B) = P(A) P(B)$. That is, the probability that two or more independent events will occur is the product of the simple probabilities. Consider, as an example, the tossing of a "fair" coin: $P(\text{head}) = 1/2$.

The probability of two heads in a row is $1/2 \times 1/2 = 1/4$, since the results of the two tosses are independent.

Consider the urn with the three balls, two white and one black, discussed above. But now suppose we replace the first ball after it is drawn. (This is known as sampling with replacement.) The draws are then independent, and the probability of two white balls in two draws is

$$P(W_1, W_2) = P(W_1) P(W_2) = 2/3 \times 2/3 = 4/9$$

EXAMPLES IN THE USE OF PROBABILITIES

Example 1—Rolling Dice

Two dice are rolled. Assuming that each die is "fair," what is the probability of rolling a seven? There are six different ways that a seven can appear. These are listed in Table 7-4.

Since the two dice are independent, the probability of obtaining a seven by any one of the ways in Table 7-4 is $1/6 \times 1/6 = 1/36$

Table 7-4
DIFFERENT WAYS OF ROLLING A SEVEN

First Die	Second Die	Probability
1	6	1/36
2	5	1/36
3	4	1/36
4	3	1/36
5	2	1/36
6	1	1/36
Total		1/6

(using the multiplication rule). The six different ways are mutually exclusive (we cannot obtain a seven two different ways at the same time). Using the addition rule, the total probability of obtaining a seven is $1/36$, taken six times $= 6/36 = 1/6$.

Example 2—Sampling

Of 50 loan accounts at a local bank, 8 are known to be behind on their payments. If 5 accounts are selected at random from the 50 accounts, what is the probability that *at least one* of the accounts selected will be behind in payments?

Note that the probability that at least 1 account selected is behind is 1 minus the probability that *all* accounts are current. So we first find the probability that *none* of the five accounts is behind (i.e., that all accounts selected are current). The probability that the first account selected is current is $P(C_1) = 42/50$. For the second account, the conditional probability of a current account, given a current account on the first selection, is $P(C_2 | C_1) = 41/49$ (of the 49 remaining accounts, 41 are current). Hence, the probability of 2 current accounts is

$$P(C_1, C_2) = P(C_1) P(C_2 | C_1) = (42/50)(41/49)$$

by use of the multiplication rule. For the third account, the conditional probability of a current account, given current accounts for the first two selected, is $P(C_3 | C_1, C_2) = 40/48$. Hence,

$$P(C_1, C_2, C_3) = P(C_1)P(C_2 | C_1)P(C_3 | C_1, C_2) = (42/50)(41/49)(40/48)$$

Continuing in this fashion, we have the probability that all 5 accounts selected are current, as

$$P(C_1, C_2, C_3, C_4, C_5) = (42/50)(41/49)(40/48)(39/47)(38/46) = 0.40$$

Then the probability that at least 1 account selected is behind is 1 minus the probability that all are current:

$$1 - 0.40 = 0.60$$

Example 3—Brand Loyalty

Marketing analysts are concerned with the loyalty of a customer to a particular brand, and with the effect of this loyalty on the brand's share of the market. There are two brands of a given product, *A* and *B*. Let us suppose that a customer who purchases Brand *A* in a given period (*t*)

has a 0.50 chance of purchasing *A* again in the next period ($t + 1$), and a 0.50 chance of purchasing Brand *B*. Those who buy Brand *B* in period t , however, have a 0.70 chance of repeating a Brand *B* purchase (they are more loyal than Brand *A* customers) and a 0.30 chance of switching to Brand *A* in period $t + 1$. This is shown in Table 7-5.

Assume that brand-buying behavior is dependent only on the immediately preceding purchase, as shown in Table 7-5, and is statistically independent of other previous purchases. Assume also that the probabilities shown in the table remain the same from period to period.

Let us suppose, at a given point in time t , that each brand has 50 percent of the market (as many customers buy *A* as buy *B*). We might

Table 7-5
PROBABILITIES OF REPEAT PURCHASES AND
BRAND SWITCHES

		Brand Purchased in Period ($t + 1$)	
		Brand <i>A</i>	Brand <i>B</i>
Brand Purchased in Period (t).....	Brand <i>A</i>	0.50	0.50
	Brand <i>B</i>	0.30	0.70

ask what will happen to the market share of each brand after one period has elapsed (time $t + 1$). During the period, Brand *A* has kept 0.50 of its own customers and captured 0.30 of Brand *B* customers. That is, the shares at time $t + 1$ are:

$$\begin{aligned}
 \text{Brand } A &= (.50)(50 \text{ percent market share of } A) + (.30)(50 \text{ percent market share of } B) \\
 &= 40 \text{ percent of the market} \\
 \text{Brand } B &= (.70)(50 \text{ percent market share of } B) + (.50)(50 \text{ percent market share of } A) \\
 &= 60 \text{ percent of the market}
 \end{aligned}$$

At the end of the first period, Brand *B* has increased its share to 60 percent of the market. The process is repeated during the second period, so that the shares at time $t + 2$ are

$$\begin{aligned}
 \text{Brand } A &= (.50)(40 \text{ percent market share of } A) + (.30)(60 \text{ percent market share of } B) \\
 &= 38 \text{ percent of the market} \\
 \text{Brand } B &= (.70)(60 \text{ percent market share of } B) + (.50)(40 \text{ percent market share of } A) \\
 &= 62 \text{ percent}
 \end{aligned}$$

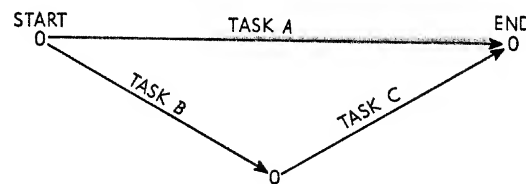
Again, Brand *B*'s share increases, but only slightly. If the process is repeated over many periods, an equilibrium is reached with Brand *A* having three eighths of the market and Brand *B* having five eighths of the market. At this point, the number of customers leaving Brand *A* is exactly balanced by those switching from *B* to *A*.

Many marketing strategies (such as pricing, advertising, and merchandise deals) are aimed at influencing brand loyalty (i.e., influencing the probabilities such as those shown in Table 7-5). The above probability analysis traces the effects of these strategies on market share.

Example 4—Project Scheduling

Construction or research and development projects require the scheduling and coordination of large numbers of tasks. It is usually important

Chart 7-2
ORDER OF TASKS



to complete the project by a scheduled date. When the times to complete some of the tasks are uncertain, the project completion time itself is uncertain. However, we can determine the probability for completion at any time.

Consider the following simplified example. A project is made up of three tasks, designated *A*, *B*, and *C*. Task *B* must be completed before *C* can start. Task *A* is not dependent upon *B* and *C* (it is done in parallel) but both *A* and *C* must be completed before the project is considered finished. This arrangement, with lines indicating tasks, is illustrated in Chart 7-2.

The time needed to complete each task is uncertain, owing to weather conditions and other unpredictable factors. However, probabilities are assigned to task completion times as shown in Table 7-6.

Let us denote the event "Task *A* takes four weeks to complete" by the symbol *A*-4. Similarly, we have *A*-6, *B*-1, etc. Assume that task completion times are independent—the time taken to complete *B*, for example, does not influence the time for *C*.

We wish to determine the probabilities associated with total project completion time. If the events *A*-4, *B*-1, and *C*-2 all occur, the total

Table 7-6

PROBABILITIES AND TIMES TO COMPLETE
TASKS A, B, AND C

Task	Completion Time, Weeks	Probability
A	4	0.50
	6	0.50
		1.00
B	1	0.25
	3	0.75
		1.00
C	2	0.80
	4	0.20
		1.00

project will take four weeks (this is the four weeks required for A; the B and C tasks take only a total of three weeks). Hence, the probability of the event $T-4$ (total project time equals four weeks) is

$$P(T-4) = P(A-4, B-1, C-2) = P(A-4)P(B-1)P(C-2) \\ = (0.50)(0.25)(0.80) = 0.10$$

using the multiplication rule for independent events.

The event $T-5$ can be obtained either by the set of events $A-4, B-1, C-4$ or by the set $A-4, B-3, C-2$. These sets are mutually exclusive—either one or the other happens, not both; and

$$P(A-4, B-1, C-4) = (0.50)(0.25)(0.20) = 0.025$$

$$P(A-4, B-3, C-2) = (0.50)(0.75)(0.80) = 0.300$$

Hence, the probability of $T-5$ is the sum: . . . 0.325

The probabilities for the values of $T-6$ and $T-7$ can be determined in a similar manner and are shown in Table 7-7.

Table 7-7

PROBABILITIES AND TIMES TO
COMPLETE TOTAL PROJECT

Project Completion Time, Weeks	Probability
4	0.10
5	0.325
6	0.425
7	0.15
	1.000

From simple probability information about the time to complete individual tasks, we have determined a complete set of probabilities for total project time.

PROBABILITY DISTRIBUTIONS

Consider an example of tossing four coins. The probabilities for various numbers of heads (r) are shown in Table 7-8 and are graphed in Chart 7-3. Note that this table simply expresses a functional relationship between values of a variable r and another set of values $P(r)$. This type of function is called a *probability distribution*. We call the

Table 7-8

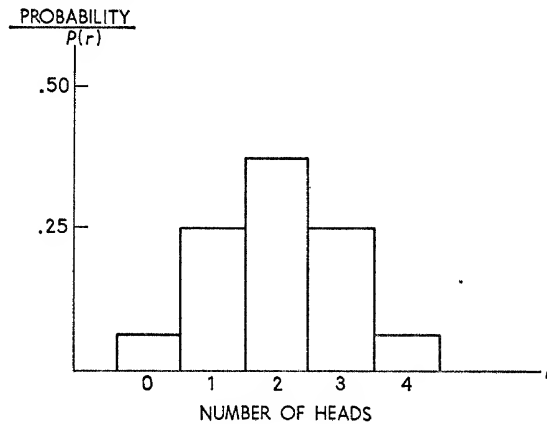
PROBABILITIES OF VARIOUS NUMBERS OF HEADS IN FOUR TOSSES OF A FAIR COIN

Number of Heads, r	Probability, $P(r)$
0	1/16
1	1/4
2	3/8
3	1/4
4	1/16
	1.0

variable r (number of heads) a *random variable*. It is random in the sense that we cannot predetermine the exact value that the variable will take on any trial; only the *probabilities* that it will take certain values are known. Each probability $P(r)$ applies to a given value of

Chart 7-3

GRAPH OF PROBABILITY FUNCTION OF TABLE 7-8



r . As noted above, each value of $P(r)$ must be between 0 and 1, and the total probabilities of mutually exclusive and collectively exhaustive events (e.g., for 0, 1, 2, 3, and 4 heads) must equal 1.

Discrete and Continuous Distributions

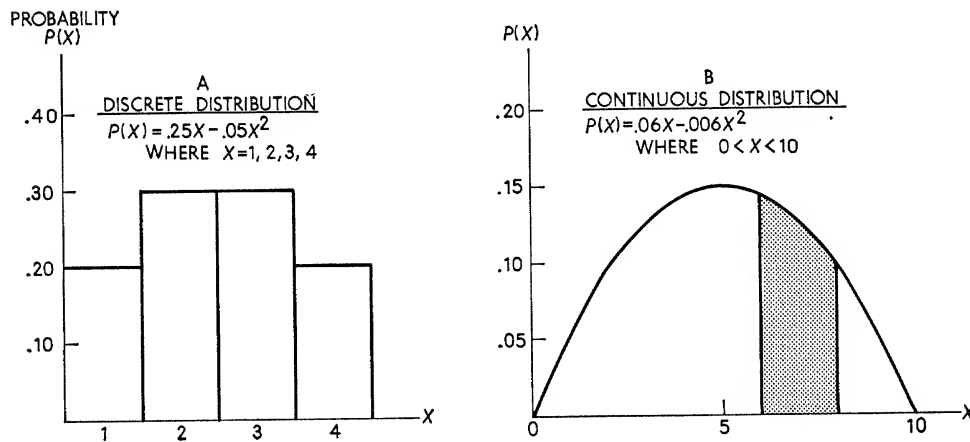
A probability distribution is *continuous* or *discrete* depending on whether the random variable can take on any real number in a specified interval or is restricted to specific values (often integers).

The distribution above is discrete, since the random variable r can take on only specific integer values. There are 0 or 1 or 2 or 3 or 4 heads in four flips of a coin. It is not possible to get $1\frac{1}{2}$ heads or 1.648 heads. On the other hand, the distribution of diameters of ball bearings is continuous since the value of the random variable can take on any value (if we have fine enough measuring instruments).

In the probability distributions in Tables 7-7 and 7-8, the relationship between the random variable and the probability function is defined by the table itself. Other probability distributions may be defined by mathematical equations. For example, the function $P(X) = 0.25X - 0.05X^2$ may define a discrete probability distribution in which the random variable X can take on the integer values 1, 2, 3, or 4. Similarly, the continuous function $P(X) = 0.06X - 0.006X^2$ may define a continuous probability distribution in which the random variable can take on any value between 0 and 10 (i.e., $0 < X < 10$). The graphs of these functions are shown in Chart 7-4. Three specific

Chart 7-4

EXAMPLES OF PROBABILITY DISTRIBUTIONS DEFINED BY MATHEMATICAL EQUATIONS



probability distributions, defined by mathematical equations, are studied in detail in Chapter 8.

Graphs of Probability Distributions

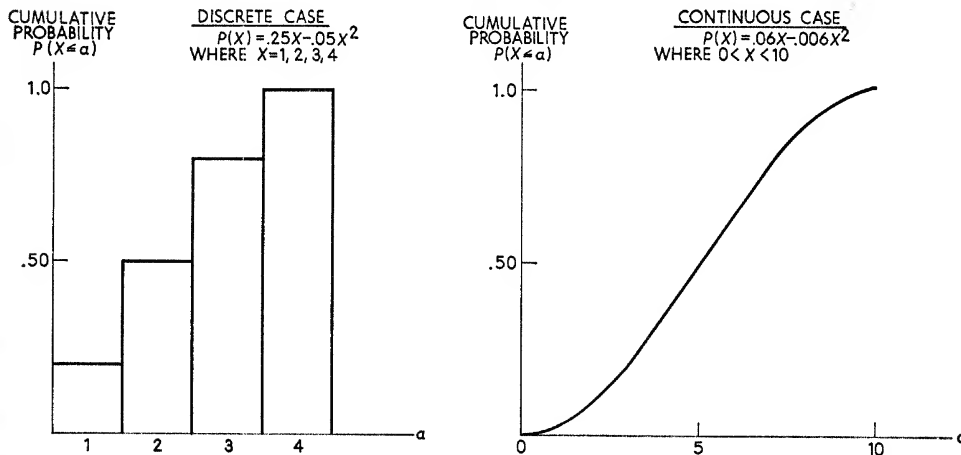
Graphs of discrete probability distributions are illustrated in Charts 7-3 and 7-4A. The values of the random variable are shown on the X axis and the associated probabilities on the Y axis. This histogram is the same as those in Chapter 4, except that the vertical scale shows probability rather than frequency.

Continuous probability distributions are represented by a smooth curve, such as in Chart 7-4B. However the values of $P(X)$ represent only the height of the curve at any point X and are *not* probabilities. In a continuous distribution, the probability of the random variable having any *exact* value is infinitely small. We can only speak of the probability of a random variable being in a specified range. For example, the probability that X falls between 6 and 8, or $P(6 < X < 8)$, is represented by the shaded area in Chart 7-4B. The total area under the curve (i.e., the probability for all values of X) is taken as 1. Thus, probability is associated with an area under the curve for continuous distributions.

It is sometimes convenient to have graphs of the probability that a random variable is less than (or greater than) a given value. These graphs of cumulative distributions are like the ogives of Chapter 4, except that probabilities are cumulated and plotted instead of frequencies.

Chart 7-5

CUMULATIVE DISTRIBUTIONS



EXPECTED VALUE AND VARIANCE OF PROBABILITY DISTRIBUTIONS

The *expected value* of a discrete random variable X is defined as follows:

$$E(X) = \sum X \cdot P(X)$$

where $P(X)$ is the probability for each value of X .

Note that we multiply each value of X by its probability and sum the

Table 7-9

PROBABILITY DISTRIBUTION OF CAR SALES
EXPECTED VALUE AND VARIANCE

Cars Sold (X)	Probability $P(X)$	$X \cdot P(X)$	$X - E(X)$	$[X - E(X)]^2$	$[X - E(X)]^2 \cdot P(X)$
0	0.20	0	-2	4	0.80
1	0.25	0.25	-1	1	0.25
2	0.25	0.50	0	0	0
3	0.10	0.30	1	1	0.10
4	0.10	0.40	2	4	0.40
5	0.05	0.25	3	9	0.45
6	0.05	0.30	4	16	0.80
Total	1.00	2.00			2.80

products. The concept of expected value then corresponds to that of a weighted mean, $\bar{X} = \sum fX/n$, where the probability $P(X)$ is equivalent to the relative frequency f , and $n = 1$, since the sum of the probabilities equals 1.

Consider a new car agency that sells from 0 to 6 cars (X) a day. In a typical period, the agency makes no sales on 20 percent of the days; it sells 1 car on 25 percent of the days, and so on, as shown in Table 7-9. These relative frequencies might be used as estimates of probabilities $P(X)$ for future sales.

To find the expected value, multiply X by $P(X)$ and sum the products (column 3):

$$E(X) = \sum X \cdot P(X) = 2.00$$

That is, average or expected sales are 2 cars a day. The expected value is called the *first moment* of a probability distribution.

The principal measure of dispersion for a probability distribution is the *variance* (the square of the standard deviation or σ^2) which is defined as:

Variance = $\Sigma[X - E(X)]^2 \cdot P(X)$ in a discrete distribution.

This is equivalent to the formula $s^2 = \Sigma fx^2/n$ (Chapter 6)² where $P(X)$ is used in place of f ; $X - E(X) = X - \bar{X} = x$, and $n = 1$. To compute the variance, take the deviation from the mean, that is, $X - E(X)$, square it, multiply it by the probability $P(X)$, and then sum the products (columns 4 to 6).

For the car sales,

Variance = 2.80 (column 6, bottom)

Standard deviation = $\sqrt{2.80} = 1.67$ cars

The variance is called the *second moment about the mean*. The further the individual values of X are from the mean, the larger the second moment.

We could define the third moment about the mean (a measure of skewness) and fourth moment (a measure of kurtosis) and so on. These, however, have limited usefulness.

The calculation of the expected value and variance for continuous distributions requires the use of the calculus. (See the appendix at end of this chapter.) However, the basic notions of the expected value as an average and the variance as a measure of dispersion apply to continuous distributions also.

SUMMARY

A *probability* is a number between 0 and 1, describing the relative likelihood of a possible event. Probabilities are often thought of as the limit of the ratio of "successes" to total trials in a long-run experiment. However, probabilities may be estimated from any of three sources: (1) the relative frequency of past events, based on either an experiment or survey; (2) theoretical distributions; or (3) the subjective judgment of the decision-maker.

A *simple probability* is the probability of the occurrence of a single event. A *joint probability* is the probability that two or more events will both occur. A *conditional probability* is the probability of the occurrence of one event, given that some other event has occurred. A *marginal probability* is the probability of the occurrence of a single event, determined as the sum of the joint probabilities involving that event.

Two events are *statistically independent* if the conditional probability of one, given the other, is equal to the simple probability of the first;

² The denominator $n - 1$ does not apply here.

that is, if $P(A | B) = P(A)$. Independence implies that knowledge of one event is of no value in predicting the other event.

If two events are mutually exclusive, the probability that one *or* the other will occur is the *sum* of the respective simple probabilities; that is, $P(A \text{ or } B) = P(A) + P(B)$. If the events are not mutually exclusive, the probability that one or the other will occur is the sum of the respective simple probabilities minus the joint probability of the two events: $P(A \text{ or } B) = P(A) + P(B) - P(A, B)$.

The joint probability that two events (A and B) will both occur is the conditional probability of one, given the other, times the simple probability of the second; that is, $P(A, B) = P(A | B)P(B)$. When the events are independent, $P(A | B) = P(A)$, so the joint probability is merely the product of the simple probabilities: $P(A, B) = P(A)P(B)$.

A probability distribution is a functional relationship between a random variable (r) and a set of probabilities $P(r)$. Probability distributions may be discrete or continuous, depending on whether the random variable can take on only a restricted set of values (e.g., only integers) or can take on any value within an interval. Probabilities may be graphed in the same way as are frequencies in Chapter 4.

The *expected value* of a discrete probability distribution is the weighted average of the random variable, the weights being the respective probabilities, that is, $E(X) = \sum X \cdot P(X)$. The *variance* of a discrete probability distribution is the sum of the deviations from the mean squared times the respective probabilities:

$$\sigma^2 = \sum \{ [X - E(X)]^2 P(X) \}.$$

The standard deviation is the square root of the variance. These general concepts will be applied to three specific probability distributions in the next chapter.

APPENDIX: EXPECTED VALUE AND VARIANCE OF CONTINUOUS DISTRIBUTIONS

Definition. A continuous distribution $f(X)$ with random variable X is a function such that

$$f(X) \geq 0 \text{ for all } X, \text{ and} \\ \int_{\text{all } X} f(X) dX = 1.0$$

Expected Value. The expected value of the random variable X is defined to be

$$E(X) = \int_{\text{all } X} Xf(X) dX$$

Thus, for the function $f(X) = 0.06X - 0.006X^2$, $0 < X < 10$,

$$\begin{aligned} E(X) &= \int_0^{10} X(0.06X - 0.006X^2) dX = \frac{0.06X^3}{3} - \frac{0.006X^4}{4} \Big|_0^{10} \\ &= 20 - 15 = 5 \end{aligned}$$

In general, the expected value of any expression involving X , say $g(X)$, is

$$E(g(X)) = \int_{\text{all } X} g(X)f(X) dX$$

Variance. The variance (σ^2) is the expected value of the function $[X - E(X)]^2$.

$$\sigma^2 = E\{[X - E(X)]^2\} = \int_{\text{all } X} [X - E(X)]^2 f(X) dX$$

In our example, $E(X) = 5.0$, and

$$\begin{aligned} \sigma^2 &= \int_0^{10} (X - 5)^2 (0.06X - 0.006X^2) dX \\ &= \int_0^{10} (X^2 - 10X + 25)(0.06X - 0.006X^2) dX \\ &= \int_0^{10} X^2(0.06X - 0.006X^2) dX \\ &\quad - 10 \int_0^{10} X(0.06X - 0.006X^2) dX \\ &\quad + 25 \int_0^{10} (0.06X - 0.006X^2) dX \\ &= \left(\frac{0.06X^4}{4} - \frac{0.006X^5}{5} \right) \Big|_0^{10} - 10(5) + 25(1) \\ &= (150 - 120) - 50 + 25 = 5.0 \end{aligned}$$

and the standard deviation $\sigma = \sqrt{5.0} = 2.24$

Evaluation of Probabilities. The integration operation can be used to measure areas under curves and hence to evaluate probabilities for continuous distributions. For example, the probability that X is between 5 and 7 in our example is

$$\begin{aligned}
 P(5 < X < 7) &= \int_5^7 (0.06X - 0.006X^2) dX \\
 &= 0.03X^2 - 0.002X^3 \Big|_5^7 \\
 &= 0.284
 \end{aligned}$$

PROBLEMS

1. An automobile dealer classified his car sales over the last year as in the following table:

Purchase of Cars and Method of Payment (Percent of Total Sales)		
Type of Car Purchased	Method of Payment	
	Cash	Credit
	New Car	Used Car
	6	18
	30	46

- In selecting a purchaser at random, what is the simple probability of new car purchase?
 - What is the joint probability of selling a used car on credit?
 - What is the conditional probability that a used car purchaser will pay cash?
 - Is the type of car purchased independent (in the statistical sense) of the method of payment? Why?
2. Suppose businessmen read periodicals as follows:

	Percent
<i>Fortune</i>	5
<i>U.S. News</i>	15
<i>Wall Street Journal</i>	15
None of the above	15
<i>Fortune</i> and <i>U.S. News</i>	5
<i>Fortune</i> and <i>Wall Street Journal</i>	15
<i>U.S. News</i> and <i>Wall Street Journal</i>	10
All three	20
Total	100

- a) If a certain businessman reads *Fortune* and the *Wall Street Journal*, what is the probability that he also reads *U.S. News*?
- b) What proportion of businessmen read *Fortune*?
- c) Are the events "reader of *Fortune*" and "reader of the *Wall Street Journal*" independent events?
- d) Are the events "reader of *U.S. News*" and "reader of the *Wall Street Journal*" independent?

3. An investor classified the stocks in his portfolio in the following manner:

	Industrial Stocks Percent	Utility Stocks Percent
Large companies (in top 100 of assets)		
Price increased (in past year)	4	1
Price decreased	8	7
Total	12	8
Small companies		
Price increased	17	3
Price decreased	55	5
Total	72	8
Total (100%)	84	16

In this portfolio:

- a) If a stock were drawn at random, what is the probability that it was one that had increased in price? What kind of probability is this? (Simple, joint, marginal, or conditional?)
 - b) What is the probability of a stock increasing if it was a large company industrial stock? What kind of probability is this?
 - c) Is size of company independent of price behavior in this portfolio? Why?
 - d) Is the type of stock (industrial versus utility) independent of the price behavior in this portfolio? Why?
 - e) Is price behavior independent of both size and type of stock? Explain.
4. Suppose 70 percent of the corporations in a certain industry have a lawyer on the board of directors and suppose 40 percent have a banker on the board. What proportion of the corporations have neither a banker nor a lawyer on the board?
5. In analyzing sales of a certain product in a retail store over the past year you discover that 10 percent of the purchases were made by men and 20 percent of the purchases were over \$10 in value. If you know that 80 percent of male customers make purchases over \$10:
- a) What percent of purchases over \$10 are made by men?
 - b) What percent of purchases are made by men or are over \$10?
6. If 30 percent of the households in a certain city have electric dryers and 40 percent have electric stoves, and if 25 percent of those who have electric stoves also have electric dryers, what proportion of those who have electric dryers also have electric stoves?

7. A market research firm is interested in surveying certain attitudes in a small community. There are 125 households broken down according to income, ownership of a telephone, and ownership of a TV.

	Households with Annual Income of \$8,000 or Less		Households with Annual Income above \$8,000	
	<i>Telephone Subscriber</i>	<i>No Telephone</i>	<i>Telephone Subscriber</i>	<i>No Telephone</i>
Own TV set	27	20	18	10
No TV set	18	10	12	10

- What is the probability of obtaining a TV owner in drawing at random?
 - If a household has income in excess of \$8,000 and is a telephone subscriber, what is the probability that it has a TV?
 - What is the conditional probability of drawing a household that owns a TV given that the household is a telephone subscriber?
 - Are the events "ownership of a TV" and "telephone subscriber" statistically independent?
 - Are the events "income of \$8,000 or less" and "ownership of TV" independent events?
8. As a bond salesman, you are considering using a list of stockholders for direct mail advertising. You know that 40 percent of investors hold stocks only and 10 percent hold bonds only, while another 20 percent hold both stocks and bonds, and the other 30 percent hold neither. Then, if an investor is a stockholder, what is the probability that he is also a bondholder?
9. A piece of electronic equipment has three essential parts. In the past Part *A* has failed 20% of the time; Part *B*, 40% of the time; and Part *C*, 30% of the time. Part *A* operates independently of Parts *B* and *C*. Parts *B* and *C* are interconnected, however, so that failure of either part affects the other. In those instances when Part *C* failed, the chances were two out of three that Part *B* would also fail.
- Assume that at least two of the three parts must operate to enable the equipment to function. What is the probability that the equipment will function?
10. If an employee shirks his work 30 percent of the time, what is the probability that he will be caught if his boss checks on him four times at random?
11. As manager at a crucial point in a ball game, you feel that your pitcher has a 70 percent chance of getting the next batter out. You could replace him with a relief pitcher who has a 90 percent chance of getting the batter out if he is at his best, but only a 40 percent chance if he is not at his best. Your pitching coach in the bullpen informs you that, on the basis of watching his warming up, he feels that the relief pitcher has about 70 percent chance of being at his best. Do you change pitchers?
12. Which of the following functions are probability distributions? Explain.
- $P(X) = X/10$ for $X = 1, 2, 3, 4$.

- b) $P(X) = X^2/10$ for $X = 1, 2, 3, 4$.
 c) $P(X) = 0.40 - 0.02X^2$ for $X = 1, 2, 3, 4$.

13. Find the expected value and variance of the distribution shown in Table 7-7.
 14. Find the expected value and variance of the distribution shown in Table 7-8.
 15. Find the expected value and variance of the probability distribution
 $P(X) = 0.25X - 0.05X^2$ for $X = 1, 2, 3, 4$.
 16. The following represents a probability distribution for the number of orchids (Z) demanded by customers in a certain florist shop:

Number Demanded Z	Probability $P(Z)$
0	0.05
1	0.10
2	0.25
3	0.30
4	0.20
5	0.10
6 and up	0
	1.00

Calculate the expected value and variance of Z .

17. Consider the probability distribution given by the following table:

X	$P(X)$
0	0.18
1	0.32
2	0.20
3	0.12
4	0.08
5	0.06
6	0.03
7	0.01
	1.00

- a) What is the expected value of X ?
 b) What is the variance of X ?
 c) What is the conditional probability that $X = 2$, given that X is an even number or zero?
 18. Consider Example 3 on page 149. Suppose the following probabilities represent the probabilities of repeat purchases or switches:

Brand Purchased in Period (t)	Brand Purchased in Period ($t + 1$)	
	Brand A	Brand B
Brand A	0.40	0.60
Brand B	0.40	0.60

Show that Brand A—40 percent, Brand B—60 percent, is an equilibrium distribution of market shares; i.e., market shares are the same in period ($t + 1$) as they are in (t).

19. Carry through the illustration of Example 4, page 151, on the assumption that there is a 0.3 probability of Task *A* taking four weeks, and a 0.7 probability of its taking six weeks.
20. A company has two warehouses, *A* and *B*. Each warehouse carries a normal stock of three units of a certain product. Daily demand (requests) for this product at *each* warehouse has the following probability distribution:

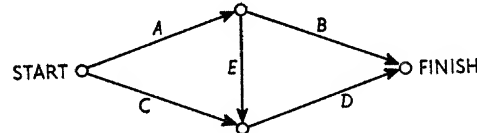
Daily Demand, Units	Probability
1	0.30
2	0.40
3	0.20
4	0.10
	1.00

- a) What is the probability that warehouse *A* will have more demand than stock on a given day?
- b) What is the probability that one or the other warehouse (not both) will have more demand than stock on a given day?
- c) What is the probability that both warehouses will have more demand than stock available on a given day?
21. Suppose that the company in Problem 20 consolidated warehouses *A* and *B* into a central warehouse *C*. A normal stock of six units is to be carried at the central warehouse *C*.
- a) Determine the probability distribution of demand for warehouse *C* from the individual distributions for *A* and *B*. [*Hint*: The probability of a demand for three units at *C* is (probability of one demand at *A* times the probability of a two demand at *B*) plus (probability of a 2 demand at *A* times the probability of a 1 demand at *B*) etc.]
- b) From the distribution determined in (a), what is the probability of having one more demanded than stock available? Of having two more demanded than stock available? Compare these with the answers to Parts (b) and (c) of Problem 20. If the answers are different, why are they so?
22. Management of the Alzo Company is considering marketing a new product. Market research indicates that there is a 0.40 probability that the total market for the product is 10,000 units; a 0.40 probability for an 8,000 unit total market; and a 0.20 probability for a 6,000 unit market.
- It is not known whether Alzo's competitor, Barden, will offer a similar product. Chances are about 50/50 that Barden will. If Barden *does not* offer a competitive product, then Alzo will capture the entire market. If Barden *does* enter the market, it will capture part of the market depending upon the price charged. If Barden sets a competitive price, Alzo management feels that Barden will have 0.20 chance of taking 60 percent of the market, a 0.50 chance of taking 40 percent of the market, and a 0.30 chance of taking 20 percent of the market. On the other hand, if Barden resorts to price-cutting, Barden has a 0.70 chance of taking 60 percent of the market and a 0.30 chance of taking 40 percent of the market.

Based upon past experience, Alzo felt the chances were 3 out of 4 that Barden would set a competitive price.

Determine the probability distribution for number of units sold. What is expected sales?

23. Suppose that in Problem 22 Barden's pricing strategy depended upon the size of the market, so that if the market was 10,000 or 8,000 units, the chances were 8/10 that Barden would set a competitive price. But if the market was only 6,000 units, the chances were 6/10 that Barden would resort to price-cutting. Determine the probability distribution for sales (units) and expected sales.
24. A project is composed of five tasks, *A*, *B*, *C*, *D*, and *E*. The order in which the tasks must be performed is shown in the network diagram (lines



represent tasks). That is, Task *A* must be done before either *B* or *E* can start; both *C* and *E* must be done before *D* can start; and both *B* and *D* must be done before the project is considered finished. Thus, there are three sequences of tasks (called paths through the network) that can hold up total project completion time: *A-B*, *C-D*, and *A-E-D*. The total project completion time is the time taken to complete the longest of these sets of tasks. For example, if *A* takes 5 weeks; *B*, 6 weeks; *E*, 2 weeks; *C*, 9 weeks; and *D*, 4 weeks; then *A-B* is 11 weeks; *C-D* is 13 weeks; and *A-E-D* is 11 weeks. The total project time is 13 weeks, determined by the *C-D* set of tasks.

The table below lists the times and probabilities to complete each of the tasks.

<i>Task</i>	<i>Time to Complete (weeks)</i>	<i>Probability</i>
<i>A</i>	5	0.50
	7	0.50
<i>B</i>	6	0.80
	9	0.20
<i>C</i>	5	0.40
	9	0.60
<i>D</i>	4	0.50
	6	0.50
<i>E</i>	2	1.00

Determine the probability distribution for project completion time. Calculate the expected completion time.

SELECTED READINGS

Selected readings for this chapter are included in the list that appears on page 188.

8. THE BINOMIAL, POISSON, AND NORMAL DISTRIBUTIONS

THIS CHAPTER describes three probability distributions that govern the behavior of many business processes. These probability distributions will be used in Chapter 9, together with the economic consequences of business actions, to develop a rational procedure for decision-making under uncertainty. In addition, the distributions will serve as a basis for evaluating sample evidence (Chapter 11).

In Chapter 4 we classified statistical data into two categories: *attributes*, which are classified into two or more discrete groups (e.g., heads or tails), and *variables*, which can be measured along a scale. The binomial and Poisson distributions describe the behavior of attributes, while the normal distribution describes the behavior of variables.

THE BINOMIAL DISTRIBUTION

We shall first discuss a few examples of the binomial distribution to illustrate the points involved. Consider the following kinds of problems:

1. What is the probability of getting 4 heads in 10 flips of a coin?
2. If a certain district is 60 percent Republican, what is the probability of getting fewer than 30 Democrats in a sample of 100 voters?
3. If a certain process produces transistors, 4 percent of which (on the average) are defective, what is the probability of getting more than 4 defectives out of 50 items?

Bent Coin Example

A coin is bent so that it turns up heads 60 percent of the time. We can ask the following question: "What is the probability of 5 heads in 5 flips?"

The events are independent; using the multiplication rule:

$$\begin{aligned}\text{Probability of 5 heads} &= P(5 \text{ heads}) \\ &= 0.6 \times 0.6 \times 0.6 \times 0.6 \times 0.6 \\ &= 0.078\end{aligned}$$

Now, what is the probability of 3 heads in 5 flips? If the order is specified (e.g., HHHTT) we can answer the question exactly as above:

$$\begin{aligned}P(3 \text{ heads in the order HHHTT}) &= 0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.4 \\ &= 0.6^3 \times 0.4^2 \\ &= 0.034\end{aligned}$$

In general, this probability is $p^r q^{(n-r)}$; the symbols being described below.

In any other order, the answer is still the same, thus:

$$\begin{aligned}P(3 \text{ heads in order TTHHH}) &= 0.4 \times 0.4 \times 0.6 \times 0.6 \times 0.6 \\ &= 0.034\end{aligned}$$

Hence, the order is unimportant, so we need to know how many ways (that is how many arrangements) 3 heads can occur in 5 flips.

This is the number of *combinations* of 5 things taken 3 at a time; that is, there are two groupings (heads and tails), and we wish to know how many ways we can arrange the 5 flips into the 2 groupings. It can be shown that the number of combinations in which r successes can occur in n trials is

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

where n factorial is $n! = 1 \times 2 \times 3 \times \dots \times n$ and $0! = 1$ by definition.

The number of combinations in which 3 heads can occur in 5 trials is, therefore,

$${}_5C_3 = \frac{5!}{3!2!} = \frac{1 \times 2 \times 3 \times 4 \times 5}{1 \times 2 \times 3 \times 1 \times 2} = 10$$

(There are 10 ways in which 3 heads can occur in 5 flips of a coin.) Let us now return to our original question (the probability of 3 heads in 5 flips of bent coin). We must multiply the number of combinations of 3

heads in 5 flips by the probability of 3 heads in 5 flips occurring in some specific order:

$$P(3 \text{ heads in 5 flips}) = 10 \times 0.034 = 0.34$$

The Binomial Probability Formula

In general, the probability of r successes in n trials is

$$P(r) = {}_nC_rp^rq^{(n-r)}$$

where r is the number of successes (i.e., heads); n is the size of the sample (i.e., number of flips); p is the probability of a success (i.e., a head); $q = (1 - p)$ is the probability of a failure (i.e., a tail); and $P(r)$ = probability of exactly r successes (i.e., r heads).

Example. Probability of 2 heads and 3 tails of our bent coin:

$$\begin{aligned} n &= 5 \text{ flips} \\ r &= 2 \text{ heads} \\ n - r &= 3 \\ p &= 0.6, \text{ the probability of a head} \\ q &= 1 - p = 0.4 \end{aligned}$$

$$P(r) = {}_nC_rp^rq^{(n-r)} = \frac{5!}{2!3!}(0.6)^2(0.4)^3 = 10 \times 0.023 = 0.23$$

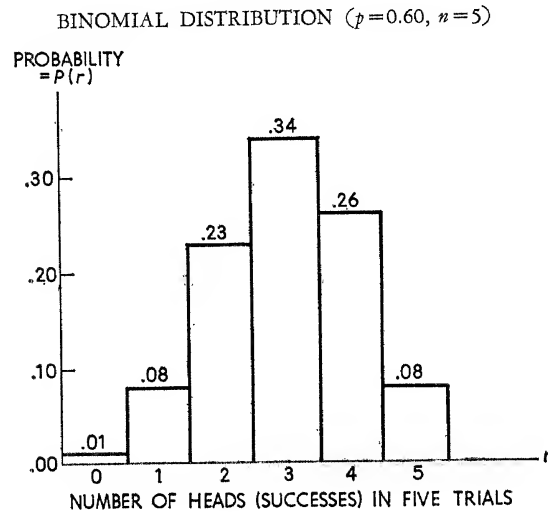
If we carried this procedure out we could find the probability of any number of heads in 5 flips of our bent coin. The results would be

Probability of 0 heads	$= P(0) = 0.01$
Probability of 1 head	$= P(1) = 0.08$
Probability of 2 heads	$= P(2) = 0.23$
Probability of 3 heads	$= P(3) = 0.34$
Probability of 4 heads	$= P(4) = 0.26$
Probability of 5 heads	$= P(5) = 0.08$
Total	<u>1.00</u>

These results can be portrayed in a histogram, plotting the random variable (heads) on the X axis and the probability on the Y axis.

This is one example of the *binomial distribution*. Note that for each flip of the coin (i.e., each trial) there were only two possible outcomes—heads or tails. We can use the same kind of analysis whenever we count only two outcomes to each trial (subject to the assumptions below), for example, when we are sampling to determine party affiliation (Democrat or Republican) or in determining if a manufactured

Chart 8-1



product is good or defective or in any case where there is only a yes or no answer.

The formula for $P(r)$ defines a whole family of distributions of r , one for each combination of the values of n and p . The quantities n and p are called the *parameters* of the binomial distribution, since they determine the probabilities for all values of r . We will use the symbol $P(r|n,p)$ to denote the probability of r given n and p .

The *expected value* or mean number of successes $E(r)$ in a binomial distribution is np , and the *variance* is npq . Thus, in the bent coin example ($n=5, p=0.60$),

$$E(r) = np = 5 \times 0.60 = 3 \text{ heads (the expected or average number of heads in 5 tosses)}$$

$$\text{Variance} = npq = 5 \times 0.60 \times 0.40 = 1.2$$

$$\text{Standard deviation} = \sqrt{1.2} = 1.1 \text{ heads}$$

Assumptions Underlying the Binomial Distribution

1. *For each trial, the random variable can take on only one of two values—success or failure.*

2. *The trials are independent.* What happens on the first trial does not affect the second, and so on. If we are flipping a coin, this means that heads will occur with the same probability, regardless of whether the previous flip was a head or a tail.

This assumption implies that we are sampling from an "infinite"

population. Flipping a coin can be considered an infinite process, for we can conceive flipping the coin forever. Likewise, if we inspect items from a lot of manufactured parts, and if *we replace each item after it is inspected*, we can again consider this an infinite population since we would never exhaust it. This latter process is called *sampling with replacement*.

Oftentimes in actual practice, we do *not* replace items in sampling from a large lot (i.e., *sampling without replacement*), and we violate the assumptions of the binomial distribution. Theoretically, we should use the *hypergeometric distribution* instead, when sampling without replacement from a finite population. This will not be described here since, in the great majority of practical applications, it can be approximated by the binomial distribution. This is because the binomial is approximately equal to the hypergeometric if the sample size (i.e., the number of trials) is small *relative* to the number of items in the population. A good rule of thumb is 20 percent. That is, if the sample size is less than 20 percent of the total number of items in the whole population, then the binomial distribution can be used even when sampling without replacement.

3. *The value of p , the probability of success, remains the same from trial to trial.* The assumption implies that, for example, the coin does not become more and more bent as the trials proceed, or that a machine does not wear and produce a higher proportion defective over time.

Mathematically, we can derive the binomial distribution from these three assumptions. If a process in the real world satisfies these assumptions, then we use the binomial probabilities to represent the real world probabilities.

Tables of the Binomial Distribution

Calculating the binomial probabilities from the formula

$$P(r) = {}_nC_rp^rq^{n-r}$$

would be quite time consuming if n were very large. Hence we resort to tables for obtaining the values. Several comprehensive tables are available.¹ We have included a shorter set of tables in Appendixes F and G at the end of the book. Appendix F lists the individual probabilities in the binomial distribution for values of n from 2 to 25, and for various

¹ See, for example, *Tables of the Binomial Probability Distribution*, U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series No. 6 (Washington, D.C.: U.S. Government Printing Office, 1949).

values of p from 0.01 to 0.50. Values for p greater than 0.50 can be read from this table by reversing the definition of "success" and "failure."

Appendix G is a table of the cumulative binomial distribution. That is, it shows the probability of r or more successes for any given value of r , and for the same values of n and p as above. Examples of the use of these tables will be given below.

Examples of the Binomial Distribution

1. A large lot of a certain manufactured part is known to contain 5 percent defective parts. If a sample of three parts is drawn at random, what is the probability that none of the parts is defective?

First let us check the binomial assumptions. The first assumption says that each part can take on only two possible values. Here we have only good or defective, so we are all right on that count.

The second assumption implies that the trials (i.e., drawings) are independent. If we were to replace each part before the next is drawn this assumption would be strictly true. However, our sample size, three items, is quite small relative to the size of the large lot, so that any error introduced on this account would be small.

The third assumption implies that the value of p remains the same as we continue to sample. Since we are sampling from a fixed lot of items which does not change, the assumption is satisfied.

Having satisfied ourselves that the binomial distribution is appropriate (or a close approximation), we proceed to calculate the required probability. In our example, $p = 0.05$, $n = 3$, and $r = 0$.

Probability of zero defectives is

$$P(r = 0) = {}_3C_0 p^0 q^3 = \frac{3!}{3!0!} (0.05)^0 (0.95)^3 = 0.857$$

2. Suppose, for our second example, we use the same circumstances as above, namely, a large lot of manufactured parts which is known to contain 5 percent defective parts. Let us now, however, take a sample of 20 items, and ask the following three questions: (a) What is the probability of *exactly* 2 defective items out of the 20 sampled? (b) What is the probability of 2 *or more* defective items? and (c) What is the probability of 2 *or less* defectives?

The evaluation of the required probabilities would involve considerable calculation, so we shall look up the values in the table instead.

a. The probability of exactly 2 defectives: This value can be found

directly in Appendix F, for $n = 20$, $p = 0.05$, and $r = 2$. The value is $P(r = 2 | n = 20, p = 0.05) = 0.189$.

b. The probability of 2 or more defectives: This value can be found directly in Appendix G, for $n = 20$, $p = 0.05$, and $r = 2$. The value is $P(r \geq 2 | n = 20, p = .05) = 0.264$.

c. The probability of 2 or less defectives: This cannot be read directly from either of our tables. Instead, we recognize the fact that the probability of 2 or less defectives plus the probability of 3 or more defectives must be 1.0. In symbols,

$$P(r \leq 2) + P(r \geq 3) = 1.0 \quad \text{or} \quad P(r \leq 2) = 1.0 - P(r \geq 3)$$

Now, the probability of three or more defectives is read easily from the table: $P(r \geq 3) = 0.075$. Hence:

$$P(r \leq 2) = 1.0 - 0.075 = 0.925$$

That is, the probability of two or less defectives is equal to 1 minus the probability of 3 or more defectives.

3. Exactly 60 percent of the workers in a certain plant belong to a union. If management drew a sample of 15 workers at random from the plant, (a) what is the probability that exactly 8 will belong to the union? (b) what is the probability that 8 or more will belong?

Again, we cannot answer these questions by direct reference to the table, since the table extends only to $P = 0.50$. Hence, we must rephrase the question as follows: 40 percent of the workers are nonunion. (a) What is the probability of obtaining exactly 7 nonunion members in the sample (i.e., 8 union members + 7 nonunion members = 15 men in sample). This is

$$P(r = 7 | p = 0.40, n = 15) = 0.177$$

The probability of 7 nonunion members is equivalent to the probability of 8 union members.

Similarly (b), the probability of 8 or more union members is equivalent to the probability of 7 or fewer nonunion members (i.e., fewer than 8). As in example 2:

$$\begin{aligned} P(r \leq 7 | n = 15, p = 0.40) &= 1.0 - P(r \geq 8 | n = 15, p = 0.40) \\ &= 1.0 - 0.213 = 0.787 \end{aligned}$$

(It is suggested that the student work several exercises to be sure he understands how to evaluate binomial probabilities.)

THE POISSON DISTRIBUTION

Another discrete distribution of some practical importance is the Poisson distribution. The Poisson is like the binomial except that we conceive of a very large number of trials and a very small probability of a success on any trial. This may best be explained by an example. If we were to inspect an enameled refrigerator door of a standard size, we might find 0 blemishes or 1 blemish or 2 blemishes or even more in a given square foot of enameling. We can count the blemish spots. It is impossible to count the number of nonblemished spots (they are practically infinite). We cannot use the binomial distribution in this case because we do not know the value of n , the total number of possible spots. Or putting it another way, the binomial is defined in terms of a particular attribute which has values 0 or 1 whereas the Poisson is defined with respect to *some unit of measurement* and there may be 0, 1, 2, 3 or more outcomes (e.g., blemishes) within a given measurement unit (e.g., a square foot of enameling). In statistical quality control, therefore, the Poisson distribution is applied to the *number of defects per unit*, whereas the binomial is applied to the number of defective units (r), as described in Chapter 25.

Formula and Assumptions of the Poisson Distribution

The Poisson probability function is

$$P(X) = \frac{e^{-m} m^X}{X!} \quad \text{for} \quad X = 0, 1, 2, \dots$$

where X is the random variable, the number of occurrences per unit of measurement; m is the mean or average number of occurrences of X per unit of measurement; and e is a constant with value of 2.718. . . .

In the example of the enameling process, the random variable X is the number of blemishes in a square foot. X is an integer, since 0, 1, 2, 3, etc. blemishes only—not 1.25—can occur in a square foot. The value m need not be an integer, since the average number of blemishes can take on any value. Note that m is the only parameter of the Poisson distribution; that is, if we know only the average, we can find the probability that any specified number of blemishes will occur.

It is curious to note that the *variance* of the Poisson distribution is equal to m . Hence, the variance equals the mean, and the standard deviation is \sqrt{m} —a very simple situation indeed!

The assumptions underlying the Poisson distribution are similar to those for the binomial.

1. Within any unit of measurement, there are a large number of possible points for an occurrence, and the probability of an occurrence in any one point is very small. Further, the random variable X must be an integer 0, 1, 2, . . . within the unit of measurement.

2. *Independence*: Any number of occurrences can happen in one unit of measurement and this will not affect the number of occurrences in other units of measurements. In our enameling example, this assumption implies that 5 blemishes in one particular square foot does not influence the probabilities for any other square foot.

3. *Stability*: The value of m (the average or mean) must remain constant. Thus about the same number of blemishes, on the average, must occur at all points of the refrigerator doors inspected.

Examples of the Poisson Distribution

1. Suppose in our example that enameling blemishes occurred on the average of 1 per square foot of refrigerator door (and the assumptions of stability and independence are valid). The probability that a square foot will have 0 blemishes is

$$P(X = 0|m = 1) = \frac{e^{-1}1^0}{0!} = 0.37$$

The probabilities of 1, 2, and 3 blemishes in a square foot are

$$P(X = 1|m = 1) = \frac{e^{-1}1^1}{1!} = 0.37$$

$$P(X = 2|m = 1) = \frac{e^{-1}1^2}{2!} = 0.18$$

$$P(X = 3|m = 1) = \frac{e^{-1}1^3}{3!} = 0.06$$

2. Consider a telephone switchboard. Suppose calls arrive at random. What would this mean? Let us look at each second of time. In most seconds there would be no calls arriving; in some seconds one call would arrive. If this were all, we could treat the process as a binomial distribution. However, in some seconds 2 or 3 or even more calls may arrive. The Poisson distribution deals with this kind of process. Note, however, that the assumption of stability would be violated if more persons, on the average, called the switchboard at certain times during the day than at other times.²

² We could treat this by breaking the day up into parts, such that m was stable over each part.

3. A certain part in a machine breaks at random. We can use the Poisson distribution to evaluate the probabilities of no breakages on a certain day, of one breakage, or two breakages, or more. Note, however, that if breakage was a function of how long the part had been in operation (i.e., wear) the assumption of stability would be violated.³

Tables of the Poisson Distribution

Appendix H at the end of the book is a table of individual probabilities of the Poisson distribution for selected values of m from 0.001 to 10.0. Appendix I is a table of the cumulative Poisson distribution. The use of these tables is very similar to that of the binomial tables. An example is given below.

A certain part breaks, on the average, twice a month. What is the probability (a) that 3 breakages will occur in a given month and (b) that 3 or more breakages will occur?

$$\begin{array}{ll} \text{(a)} & P(X = 3 | m = 2) = 0.180 \quad \text{Appendix H} \\ \text{(b)} & P(X \geq 3 | m = 2) = 0.323 \quad \text{Appendix I} \end{array}$$

Poisson Approximation to the Binomial

Another important use of the Poisson distribution is as an approximation to the binomial. Indeed, we can think of the Poisson as the limiting distribution to the binomial as n becomes large and p becomes small. Thus, when n is large and p is small, we can use the Poisson to evaluate binomial probabilities.

How large must n be and how small must p be? As a rule of thumb, we can use the Poisson to approximate the binomial if

$$\begin{array}{lll} n \geq 10 \text{ and } p \leq 0.01 & \text{or} & n \geq 20 \text{ and } p \leq 0.03 \quad \text{or} \\ n \geq 50 \text{ and } p \leq 0.05 & \text{or} & n \geq 100 \text{ and } p \leq 0.08. \end{array}$$

These requirements achieve a moderate degree of accuracy in the approximation. If very fine precision is required, larger sample sizes would be required.

To approximate a binomial probability, we simply set $np = m$ and look the values up in the Poisson table.

As an example: Suppose we are sampling 1,000 items that have 0.001 fraction defective on the average, that is, $n = 1,000$, $p = 0.001$, and $np = m = 1.0$ (an average of one defective per 1,000).

³ If there are many parts, even though the life of each is a function of wear, the breakage rate of the aggregate often may be described by a Poisson distribution.

We can then estimate the probability of getting any number of defects in our sample by using the Poisson table, as follows:

$$\begin{aligned}P(0 \text{ defectives}) &= 0.368 \\P(1 \text{ defective}) &= 0.368 \text{ etc.}\end{aligned}$$

THE NORMAL DISTRIBUTION

By far the most important distribution in statistics is the normal distribution. This function was described in Chapter 4 as a continuous distribution represented by a symmetrical, bell-shaped curve (see Charts 4-5, 4-6, 6-1, and 6-2). It is useful for two purposes:

1. It portrays the distribution of a population of certain types of measurements, such as heights of men, test scores, or the prices of laying mash in Chart 4-5.

2. More important, it describes how certain measures, such as the mean, vary from sample to sample because of *chance*, that is, the normal curve portrays the frequency distribution of all possible means of large samples that might be drawn from almost any kind of population. In Chapter 11 we will show how a distribution of sample means follows this pattern, so that we can estimate the *sampling error*.

The equation for the normal distribution is

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}}$$

where X is the random variable and μ and σ are the parameters. The constant π is 3.14159 . . . and e is 2.718 For the normal distribution, the expected value or mean is $E(X) = \mu$ and the variance is σ^2 . Normal distributions can take on many different shapes, depending on the values of these two parameters. Consider, for example, Chart 6-1, panels 1 and 2. Since the normal curve is a continuous distribution, the random variable X can take on any value, rather than only discrete values, as in the binomial and Poisson distributions.

It would be difficult to measure the probabilities under the normal curve were it not for a simple transformation which makes it possible to use only a single table. The trick is simply that we discuss normal distributions and associated probabilities in terms of standard deviation (σ) units from the mean (μ) of the distribution.

It was pointed out in Chart 6-2 that in a normal distribution

$\mu \pm \sigma$	includes 68.27 percent of the values
$\mu \pm 2\sigma$	includes 95.45 percent of the values
$\mu \pm 3\sigma$	includes 99.73 percent of the values

That is, if we draw a *single* item from this distribution, the *probability* is 0.6827 (about two chances out of three) that it will fall within the interval $\mu \pm \sigma$; the probability is 0.9545 that it will fall within the interval $\mu \pm 2\sigma$, and so on. These probabilities hold for all normal distributions, regardless of the mean or standard deviation. Furthermore, we can evaluate, similarly, probabilities for any number of standard deviations difference from the mean.

Table of Areas under the Normal Curve

We can determine these probabilities from a table of areas under the normal curve. Appendix D shows the proportion of the total area which lies between the mean and any other point X along the horizontal axis. To use the table, first take $X - \mu$ and divide by σ as follows:

$$u = \frac{X - \mu}{\sigma}$$

The value u is called the *standard normal deviate* and represents the number of standard deviation units the random variable X is above or below the mean. The whole table then represents a *standardized normal distribution* with mean $\mu = 0$ and standard deviation $\sigma = 1$.

The left-hand stub and the heading of Appendix D show the values of these deviations (u) from 0.0 (the mean itself) to 5.0, a point far out under the tail of the normal curve. The body of the table shows the proportion of the total area between the mean and any given value of u . Since the normal curve is symmetrical about the mean, the table can be used for points on either side of the mean.⁴

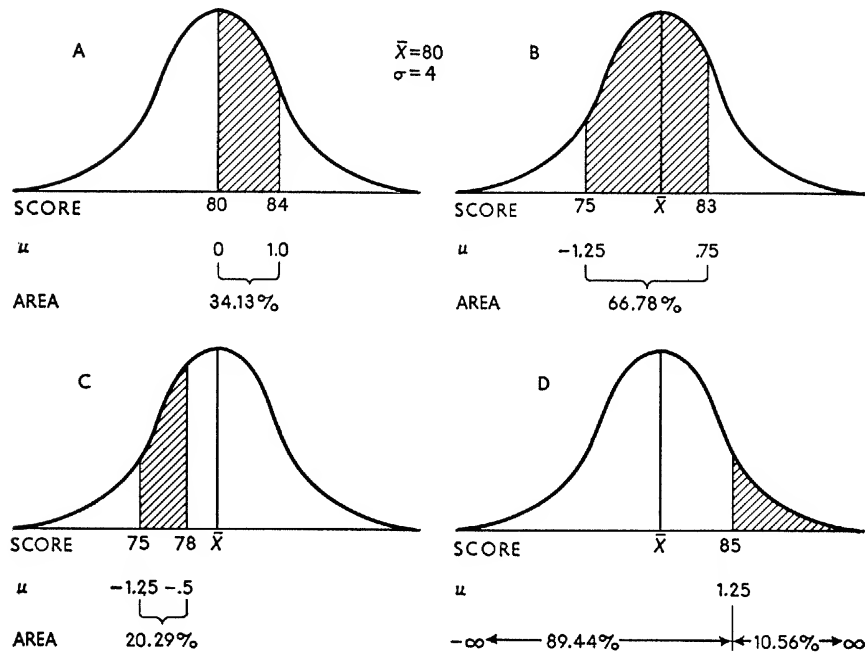
To illustrate, suppose a large number of job applicants take an aptitude test given by the personnel department of a company. The scores on the test form a normal distribution⁵ with an arithmetic mean of 80 and standard deviation of 4. Now consider the following cases. These are illustrated in Chart 8-2, panels A to D, respectively.

A. What proportion of applicants should score between 80 and 84? The deviation of the point 84 from the mean 80 is 4, so in standard deviation units, $u = 4/4 = 1.0$. Looking in Appendix D opposite $u = 1.0$, the proportion of the total area in this segment is 0.3413, or 34.13 percent. The table shows probabilities, while the chart shows relative areas. The two are equivalent, since the area under any segment

⁴ Theoretically, the curve extends indefinitely on each side of the mean without touching the X axis. However, only a negligible part of the area lies more than four or five standard deviations from the mean, so the infinite tails can be ignored.

⁵ The distribution of scores may be treated as continuous, since differences between successive scores are small.

Chart 8-2

FINDING AREA UNDER A NORMAL CURVE
IN APPENDIX D

of the curve is proportional to the probability. The proportion of scores that fall between the mean and one standard deviation on *both* sides of the mean is twice 34.13 percent, or 68.26 percent—the same value that was given for $\mu \pm \sigma$ previously (except for a slight error in rounding).

Many intervals do not terminate at the mean. These may be broken down, however, into intervals that do terminate at the mean, as shown below. Hence, Appendix D can be used for any interval.

B. What proportion of scores should fall between 75 and 83? Since these points fall on both sides of the mean, the areas between the mean and each point must be added. For the score 83, $u = (83 - 80)/4 = 0.75$. In Appendix D, look down the u column to 0.7 and across to the column headed 0.05; the area is 0.2734. Similarly, for 75, $u = (75 - 80)/4 = -1.25$, and the area is 0.3944. The combined area is then $0.2734 + 0.3944 = 0.6678$ or 66.78 percent.

C. What proportion of scores should fall between 75 and 78? Since both points are on the same side of the mean, the areas between each point and the mean must be subtracted to get the area between them. For 75, the area is 0.3944 as above. For 78, $u = -0.5$ and the area is

0.1915. The area between 75 and 78 is then $0.3944 - 0.1915 = 0.2029$, or 20.29 percent of the total area.

D. What proportion of scores should *exceed* 85? This is 50 percent—the entire segment above the mean—minus the proportion of scores between the mean and 85, or 39.44 percent (for $u = 1.25$). The answer is then 10.56 percent. Similarly, the proportion of scores *below* 85 (the unshaded part of panel D) is $50 + 39.44 = 89.44$ percent.

The table of areas under a normal curve thus serves to show the probabilities for any segment of the curve. When in doubt as to how to apply this table, draw a rough diagram, as in Chart 8-2, to picture the areas needed.

Normal Approximation to the Binomial

We noted before that when n is large and p is near 0 or 1 we can use the Poisson distribution to approximate the binomial. On the other hand, when n is large and p is *not* close to 0 or 1 we can use the *normal* distribution to approximate the binomial. How large must n be and how large must p be?

The influence of sample size and value of p on the shape of the distribution is illustrated in Chart 8-3. The chart represents the distributions of r , the number of “successes” for various combinations of values of n and p . The polygons show that the distribution of r is discrete rather than continuous. They also show how skewness depends on n , the size of sample and the population value of the proportion p .

Effect of p on the Distribution. In panel A of Chart 8-3, probability distributions of number of successes are shown for samples of a fixed size— $n = 10$ —but for varying values of p from 0.05 to 0.5. When $p = 0.05$, the distribution has a high degree of positive skewness. As the value of p approaches one half (0.5), the skewness approaches zero, so that when $p = 0.5$ the distribution is perfectly symmetrical and nearly normal.

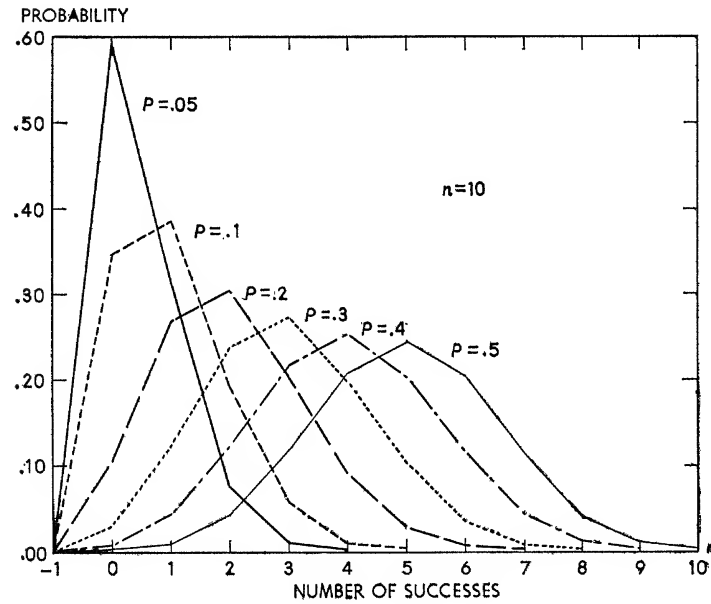
Effect of Sample Size. In panel B of Chart 8-3, probability distributions are shown for a fixed value of a proportion ($p = 0.1$), but for varying sizes of sample from 10 to 100. For small values of n the skewness is large and positive; as n increases, the approach to the symmetrical normal curve is rather striking. The same curves apply to q as for p , substituting “number of failures” for “number of successes.”

The curves illustrate the fact that n should be large, or else p should be not too close to 0 or 1, to justify the use of the methods presented below, since they are based on the assumption that the distribution of the number of successes is approximately normal. As a rule of thumb,

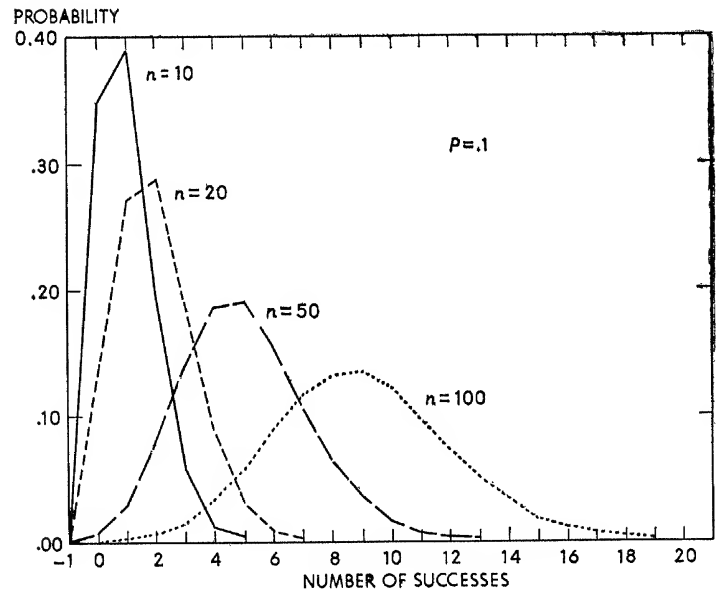
Chart 8-3

PROBABILITY DISTRIBUTIONS OF NUMBER OF SUCCESSES

A. Fixed Size of Sample, $n = 10$,
and Different Values of p



B. Fixed Value of Proportion, $p = 0.1$,
and Different Sizes of Sample



both np and nq should be about 5 or more for this assumption to be valid. Thus, if $n = 10$, p would have to be 0.5 to make $np = 5$, as in the right-hand curve of panel A. On the other hand, if $p = 0.1$, n would have to be as large as 50 (panel B) for the distribution to be roughly normal. The assumption of normality is useful both because it is valid for most practical problems involving large samples and because it is simpler than using the binomial distribution.

How can we make the approximation? Proceed as follows:

1. Set np equal to μ and \sqrt{npq} equal to σ .
2. Remember that the binomial is a discrete distribution. To allow for this we have to use a factor of $+\frac{1}{2}$ or $-\frac{1}{2}$ added to X , depending upon the circumstances. To find the probability of r or *less* successes, *add* $\frac{1}{2}$ to the value of X in calculating the normal deviate u ; to find the probability of r or *more* successes, *subtract* $\frac{1}{2}$ from the value of X in determining u .
3. Look up the probabilities in the normal table (Appendix D).

Example. The probability of a defective item is $p = 0.20$. We take a sample of 400 items from a very large lot.

- a. What is the probability of 90 or more defectives?

$$\begin{aligned}\mu &= np = 80 \\ \sigma &= \sqrt{npq} = \sqrt{400 \times 0.2 \times 0.8} = 8\end{aligned}$$

Now, the dividing line between 90 or more and the rest of the distribution is $89\frac{1}{2}$. That is, the probability of being greater than $89\frac{1}{2}$ for the continuous normal distribution is approximately the same as the probability of 90 or more in the discrete binomial.

$$u = \frac{X - \mu}{\sigma} = \frac{89\frac{1}{2} - 80}{8} = 1.19$$

$$P(u > 1.19) = 0.1170$$

- b. What is the probability of exactly 90 defectives? The probability of *more* than 90 defectives in the binomial distribution is equivalent to the probability of more than $90\frac{1}{2}$ defectives in the normal distribution. For $X = 90\frac{1}{2}$,

$$u = \frac{90\frac{1}{2} - 80}{8} = 1.31$$

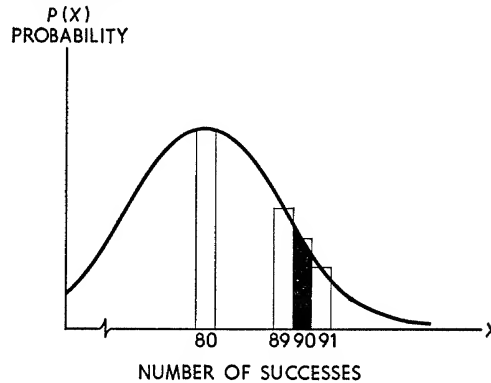
$$P(u > 1.31) = 0.0951$$

$$P(\text{exactly } 90) = P(1.19 < u < 1.31) = 0.1170 - 0.0951 = 0.0219$$

The shaded area in Chart 8-4 illustrates this probability.

Chart 8-4

NORMAL APPROXIMATION TO BINOMIAL DISTRIBUTION

**Normal Probability Paper**

Normal probability paper is special graph paper with a scale such that the *cumulative* normal distribution plots as a straight line (see Chart 8-5).

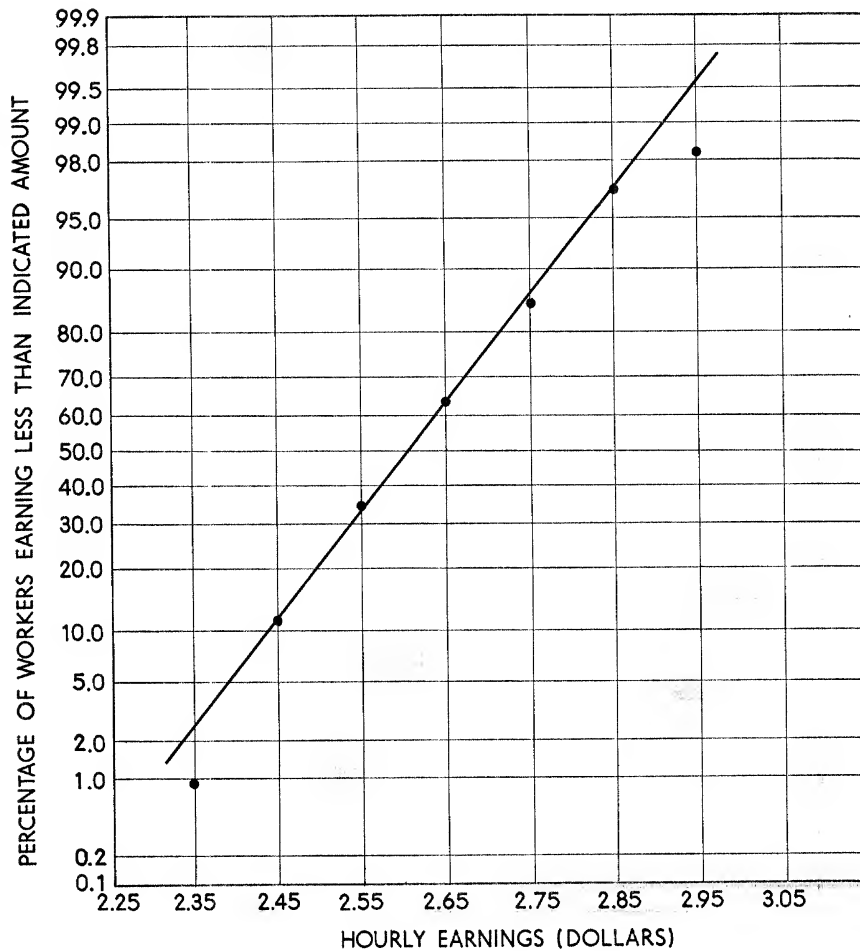
The major use of this paper is in testing whether a particular distribution is normal. For example, you have samples from some population (e.g., scores of employees on a manual dexterity test) and you wish to know if the distribution is normal. Simply plot the cumulative distribution on normal probability paper. If the distribution is normal, the points should lie close to a straight line (there will be some chance variation about the line).

The cumulative hourly earnings of 214 apprentice machine tool operators (see Table 4-6) is plotted on normal probability paper in Chart 8-5. A straight line has been drawn by inspection through these points. The five points between \$2.45 and \$2.85 lie nearly on the line, indicating that the distribution of earnings is roughly normal over this middle range. The two end points, however, are out of line; hence, the distribution is not normal near its extremes.

A second purpose of normal probability paper is to fit a normal curve to a set of sample data drawn from a normal population in order to estimate the distribution of the population. Thus, if we read the ordinates from the straight line in Chart 8-5, we can estimate the percents of *all* apprentice machine tool operators earning less than the indicated values of X . This device irons out sampling errors. For example, 64 percent of workers in the sample earned less than \$2.65 an hour, but we estimate that only 62 percent of all workers fall in this group (assuming a representative sample from a normal population of earnings).

Chart 8-5

CUMULATIVE HOURLY EARNINGS OF 214 APPRENTICE
MACHINE TOOL OPERATORS PLOTTED AS PERCENT OF
TOTAL ON NORMAL PROBABILITY PAPER



SUMMARY

This chapter describes three specific probability distributions: the binomial, the Poisson, and the normal.

The *binomial distribution* characterizes situations in which we are sampling from a population of attributes having only two values (yes or no, success or failure, etc.). It describes the number of successes (r) achieved in a fixed number of trials (n). The binomial is a discrete distribution.

The assumptions underlying the binomial are: (1) the random variable can take on only one of two values—success or failure; (2) the trials are independent; and (3) the probability of a success remains the same from trial to trial.

The *Poisson distribution*, like the binomial, is a discrete distribution. The random variable X can take on the value of 0 or any positive integer. The Poisson distribution is used to represent random occurrences in some unit of measurement—such as the number of telephone calls per unit of time or the number of defects per foot of wire.

The assumptions underlying the Poisson distribution are: (1) there is a very large number of possible occurrences in any unit of measurement; (2) there is independence from one unit of measurement to another; and (3) the average number of occurrences per unit remains the same.

If the number of trials (n) is large and the probability of success (p) is small, the Poisson distribution is a close approximation to the binomial.

The *normal distribution* is a continuous distribution represented by the familiar bell-shaped curve. The standardized normal distribution has a mean of zero and a standard deviation of one. Using this standard distribution and Appendix D, we can evaluate probabilities for any normal distribution.

If the number of trials (n) is large and the probability of success (p) is not close to either 0 or 1, the normal distribution is a good approximation to the binomial.

Normal probability paper may be used to test if a given set of data follow the normal distribution, or to estimate the distribution of a normal population from sample data.

The three distributions studied in this chapter, together with their parameters, means, variances, and standard deviations are shown in the table.

<i>Distribution</i>	<i>Parameters</i>	<i>Mean</i>	<i>Variance</i>	<i>Standard Deviation</i>
Binomial	n, p	np	npq	\sqrt{npq}
Poisson	m	m	m	\sqrt{m}
Normal	μ, σ	μ	σ^2	σ

PROBLEMS

In problems 1 through 5 below, evaluate the binomial probabilities by using the binomial probability formula.

1. What is the probability of three heads in four flips of a fair coin?

2. What is the probability of drawing (with replacement) two red chips and one yellow chip from a bag of chips containing 20 percent red and 80 percent yellow chips?
3. What is the probability of drawing three aces out of five cards from a deck of cards in which the card drawn is replaced and the deck shuffled before each draw?
4. What is the probability of drawing four successive defective parts from a large lot which is known to contain exactly 10 percent defective parts?
5. If 60 percent of television viewers are watching a certain program, what is the probability that more than half of those selected in a random sample of five will be watching the specified program?
6. Evaluate the following binomial probabilities, using Appendixes *F* and *G*.

a) $P(r = 6 n = 15, p = 0.35)$	f) $P(r \geq 9 n = 18, p = 0.60)$
b) $P(r \geq 5 n = 12, p = 0.25)$	g) $P(r < 6 n = 14, p = 0.70)$
c) $P(r < 11 n = 20, p = 0.45)$	h) $P(5 \leq r \leq 13 n = 20, p = 0.40)$
d) $P(r \leq 2 n = 16, p = 0.06)$	i) $P(1 < r < 5 n = 20, p = 0.12)$
e) $P(r = 18 n = 20, p = 0.95)$	
7. Evaluate the following binomial probabilities, using Appendixes *F* and *G*.

a) $P(r = 1 n = 8, p = 0.01)$	f) $P(r \geq 12 n = 20, p = 0.75)$
b) $P(r \geq 2 n = 13, p = 0.15)$	g) $P(r < 5 n = 15, p = 0.60)$
c) $P(r < 15 n = 20, p = 0.50)$	h) $P(7 \leq r \leq 10 n = 24, p = 0.55)$
d) $P(r \leq 6 n = 20, p = 0.20)$	i) $P(2 < r < 5 n = 18, p = 0.30)$
e) $P(r = 15 n = 25, p = 0.70)$	
8. Evaluate the following Poisson probabilities, using Appendixes *H* and *I*.

a) $P(X = 2 m = 0.20)$	c) $P(X < 5 m = 5.0)$
b) $P(X \geq 3 m = 0.80)$	d) $P(2 < X \leq 6 m = 2.4)$
9. Evaluate the following Poisson probabilities, using Appendixes *H* and *I*.

a) $P(X = 4 m = 2.6)$	c) $P(X < 2 m = 1.0)$
b) $P(X \geq 1 m = 0.40)$	d) $P(10 \geq X \geq 5 m = 6.5)$
10. A part to a certain machine is known to break randomly on the average of once in five days. How many parts must be available so that there is less than once chance in 100 of having more breakages than parts available on a given day?
11. Ships are known to arrive randomly at a port on an average of two days apart. What is the probability of two or more ships arriving on the same day?

12. The Speedo Computer averages 0.05 breakdowns requiring service per hour of operating time. What is the probability of no breakdowns in an 8-hour day? In a 40-hour week? Assume a Poisson distribution for breakdowns.
13. The random variable X is normally distributed with mean 50 and standard deviation 20. Evaluate the following probabilities:
- | | |
|-------------------|---------------------------|
| a) $P(X \geq 75)$ | c) $P(25 \leq X \leq 45)$ |
| b) $P(X \leq 55)$ | d) $P(35 \leq X \leq 80)$ |
14. The random variable X is normally distributed with mean 18 and standard deviation 10. Evaluate the following probabilities:
- | | |
|-------------------|---------------------------|
| a) $P(X \geq 28)$ | c) $P(12 \leq X \leq 16)$ |
| b) $P(X \leq 17)$ | d) $P(15 \leq X \leq 24)$ |
15. Suppose the haddock catch in Boston over the past 10 years has averaged 100 million pounds annually, with a standard deviation of 5 million pounds. For Gloucester over the same period, the mean has been 10 million pounds, with a standard deviation of 2 million pounds. If in one year the Boston catch is 108 million pounds, how large must the Gloucester catch be that year to be just as exceptional? (Assume normal distributions.)
16. The average grade on an examination taken by a large number of students is 80. The standard deviation of the grades is 6. The instructor wishes to award A's to 10 percent of the class. Assuming grades are approximately normally distributed, above what numerical grade would he give an A?
17. A firm estimates that 3 percent of its accounts receivable cannot be collected. What is the probability that out of its 200 current accounts receivable, eight or more will be uncollectable?
18. A sales manager believes that 60 percent of consumers prefer his product over his competitor's. Under this assumption, what is the probability of obtaining fewer than 54 who prefer his product out of a random sample of 100 consumers?
19. The number of misprints on a page of a daily newspaper has a Poisson distribution. You are told that the average number of misprints is $1\frac{1}{2}$ per page. You examine three pages at random and find no misprints. What is the probability of this sample result?
20. Daily demand for orchids at Joe's flower stand is approximately normally distributed with mean sales of 12 per day and standard deviation of 4 orchids. How many orchids must be on hand in the morning to assure no more than one chance in 5 of running out of orchids during the day?
21. In a recent survey, 85 of 100 firms surveyed reported an increase in sales over the same month last year. If in fact 80 percent of all firms had such a
-

sales increase, what is the probability of obtaining exactly the sample result observed? What is the probability of 85 *or more* firms out of 100 reporting sales increases?

22. The charge accounts at a certain department store have an average balance of \$120 and a standard deviation of \$40. Assuming that the account balances are normally distributed:
- What proportion of the accounts is over \$150?
 - What proportion of the accounts is between \$100 and \$150?
 - What proportion of the accounts is between \$60 and \$90?
23. The World Series is to be played between two teams, the Nationals and the Americans. Suppose that the Nationals have a superior team so that the probability of their winning in any single game is 0.60. Assume that this probability remains the same from game to game and that games are statistically independent.
- What is the probability that the Nationals will win the series (i.e., will win the necessary four games)?
 - What is the probability of the Nationals winning in four games?
 - What is the probability of the series going exactly five games and the Nationals winning?
 - What is the probability of a seven-game series (the maximum possible number)?
24. A company purchases large lots of a certain electronic component. The decision to accept these purchased lots or to reject them (return them to the supplier) is based upon a sample of 20 items. If any of the 20 items are defective, the lot is rejected; otherwise, it is accepted.
- What is the probability of rejecting a lot that has 1 percent defectives? What is the probability of accepting such a lot?
 - What is the probability of accepting a lot containing 10 percent defectives?
25. Suppose that the company in Problem 24 was considering using a sample of 50 items rather than the 20 items used previously. Assuming that a lot is accepted if fewer than 2 defectives are found and rejected if 2 or more defectives are found in the sample:
- What is the probability of rejecting a lot with 1 percent defectives?
 - What is the probability of accepting a lot with 10 percent defectives? (*Hint*: Use Poisson approximation to the binomial.)
26. Calculate the probabilities of accepting a lot for each of the sampling plans in problems 24 and 25 for the intermediate values of 0.02, 0.05, and 0.08 for the fraction defective in the lot. Plot these values and the ones calculated in Problems 24 and 25 on a chart. (The Y axis is the probability of accepting the lot; the X axis is the fraction defective in the lot). Connect the points for each plan by a smooth curve. These are the *operating characteristic curves* (or OC) for each sampling plan. Use the OC curves to compare the two sampling plans.
-

27. An auditor wishes to determine a rule to use in evaluating the accounts payable of a certain firm. There are 5,000 such accounts. The auditor considers the accounts as satisfactory if there are mistakes in only 1 percent of them. On the other hand, if 5 percent or more are in error, the auditor would require a thorough investigation. Since there are a large number of accounts, the auditor plans to take a sample of 25 accounts and investigate these. His decision to certify the accounts payable or to require further investigation will depend upon the outcome of the sample. The auditor decides to certify the accounts if none or only one account of the 25 sampled is found in error and to require further investigation if two or more accounts prove in error.
- If, in fact, 50 accounts are in error, what is the probability that the auditor will certify the accounts? What is the probability that he will decide upon further investigation?
 - If, in fact, 250 accounts are in error, what is the probability that the auditor will require further investigation? What is the probability that he will certify the accounts?

SELECTED READINGS

GOLDBERG, SAMUEL. *Probability, An Introduction*. Englewood Cliffs, New Jersey: Prentice-Hall, 1960.

A detailed and systematic treatment of discrete probability.

GUENTHER, WILLIAM C. *Concepts of Statistical Inference*. New York: McGraw-Hill, 1965.

Chapter 2 covers a variety of probability models.

HUFF, DARRELL. *How to Take a Chance*. New York: W. W. Norton, 1959.

A short and readable book dealing with probability with a humorous approach.

LEVINSON, HORACE C. *Chance, Luck, and Statistics*. New York: Dover Publications, 1963. (Paper).

Part I of this book deals with probability in an easily understandable way.

MOSTELLER, FREDERICK; ROURKE, ROBERT E.; and THOMAS, GEORGE B., JR. *Probability and Statistics*. Reading, Massachusetts: Addison-Wesley, 1961.

A detailed treatment of probability at the elementary level.

NATIONAL BUREAU OF STANDARDS. *Tables of the Binomial Probability Distribution*. Washington, D.C.: U.S. Government Printing Office, Applied Mathematics Series No. 6, 1949.

Detailed tables of the binomial distribution.

PRATT, JOHN W.; RAIFFA, HOWARD; and SCHLAIFER, ROBERT. *Introduction to Statistical Decision Theory* (Preliminary Edition). New York: McGraw-Hill, 1965.

Chapters 2 through 6 deal with the foundations of subjective probability for decision-making purposes.

SCHLAIFER, ROBERT. *Probability and Statistics for Business Decisions*. New York: McGraw-Hill, 1959.

Chapters 1, 3, and 9 discuss basic probability definitions. Chapters 10 and 13 deal with the binomial and Poisson probability distributions, respectively. Chapters 17 and 18 treat the normal distribution.

9. PROBABILITIES AND DECISION-MAKING

THIS CHAPTER combines probabilities with the economic consequences of future events, and thus formulates a logical procedure for making decisions.

CERTAINTY VERSUS UNCERTAINTY

In some business decisions, all the facts relevant to the decision are known in advance; that is, there is no uncertainty about future costs or profits. The decision problem is to select the best of the known alternatives. The "transportation problem" is an example of this type of decision situation: A firm has several factories that ship goods to its warehouses. The factories and warehouses are scattered geographically around the country. The shipping costs from each factory to each warehouse are known with certainty. The capacities of the factories and the requirements of the warehouses are also known in advance. Despite the fact that all this information is known without error, the determination of the optimum (least cost) shipping schedule (i.e., which factories should ship to which warehouses) is not a trivial problem and often requires complex mathematical techniques.¹ Note again that all relevant information is known in advance; the solution to the problem involves a search through all alternatives to find the optimum one. These are the characteristics of *decision-making under certainty*.

Contrast the problem faced by the buyer for a department store to the

¹ This is the transportation problem in linear programming. For a discussion of this problem, refer to Daniel Teichroew, *Introduction to Management Science: Deterministic Models* (New York: John Wiley, 1964) or another text on operations research or linear programming.

above illustration. The buyer must purchase in advance the merchandise needed by his store for a particular season. The cost of the merchandise and the price at which it will be sold may be known. The amount to order is what must be decided. If he orders too much merchandise, it may have to be sold at clearance prices, thereby reducing the profit for the store. Similarly, if too little merchandise is ordered, sales may be lost and the opportunity for additional profits may be forgone. To make this decision, the buyer must estimate the future demand for merchandise. Generally, he cannot know this beforehand; there is some uncertainty about the demand that will materialize owing to the appeal of the particular products, the trends in style, general economic conditions, and other factors. The buying decision is thus a *decision under uncertainty*. Such decisions are characterized by the fact that the value or one or more variables is not known to the decision-maker at the time the decision is to be made. This is not to say that no information about the value of the uncertain variable is known. The department store buyer certainly has some estimate of future demand based upon his past experience, his evaluation of the merchandise, and his knowledge of economic conditions. Therefore, he may feel that certain levels of demand are *more likely* than others.

PROBABILITY MODELS

Models or artificial representations of reality have long been useful in scientific analysis. Engineers build scale replicas of aircraft and test them in wind tunnels, or construct replicas of dams before deciding to build them. Often, an equation may be used to represent some phase of reality, as with the laws in physics. For example, the equation

$$d = \frac{1}{2} g t^2$$

predicts the distance (d) that a freely falling object will travel as a function of the time (t) it has been falling. (The g is a constant.) This model is a very useful one for describing a particular aspect of the real world.

In business decision-making under uncertainty, it is useful to use models or representatives of reality based upon probabilities and probability distributions. For example, a manufacturer may have a production process that turns out parts classified as good or defective. The binomial probability distribution may serve as a model for this process if the assumptions of this distribution are approximately satisfied.

Rarely does a model conform exactly with reality—it would have to include far too many factors and be very complex. For example, the

physical law described above does not include the air resistance to the falling object. It is unlikely that any manufacturing process exactly satisfies the binomial assumptions. However, for a model to be useful, it need represent only the important variables affecting the decision at hand. Thus, the binomial model might adequately represent the manufacturer's production process for decision-making about the quality of outgoing products.

HOW TO MAKE A DECISION

How can we use probability and probability models in making business decisions? Bear in mind that we are concerned with decision-making under uncertainty. In order to make a decision, there must be two or more possible actions or alternatives available to the decision-maker. Otherwise, there is no decision problem. And, since we are operating under uncertainty, there must be two or more events or values that can be taken on by the unknown variable. Such possible events are sometimes called *states of the world*, since they represent different happenings that can occur. The decision-maker is uncertain because he does not know which event will happen (i.e., which state of the world will materialize).

Consider the concepts in the following example. The Zip Car Rental Company rents cars at a rate of \$10 per day. (The customer pays for his own gasoline and oil.) Cars are rented for one day only. Zip Company does not own its own cars but leases them on a daily basis from a large leasing firm. The larger firm pays the maintenance cost for the cars. Zip must specify the number of cars it intends to lease on a given day at least one week in advance. The daily lease fee paid to the leasing firm by Zip Company is \$7 per day. (To avoid confusion, note that the word "lease" is used to denote the arrangement between Zip Company and the large leasing firm; the words "rent" and "rental" are used to denote relationships between Zip Company and its customers.)

Zip is faced with the decision of how many cars to lease for a given day one week hence. The demand for rental cars varies from day to day. If Zip Company leases more cars than are requested as rentals on a particular day, Zip Company will lose the lease fee of \$7 for each car unrented. If demand for cars is greater than the number available, a profit of \$3 per car (the \$10 rent less the \$7 lease fee) is forgone.

In this decision situation, the unknown factor is the number of rental requests for a given day. The possible happenings, or *states of the world*, are thus the events: "10 requests for rental cars"; "11 requests for rental cars"; "12 requests"; etc. The actions or alternatives available to the

decision-maker are: "lease 10 cars"; "lease 11 cars"; etc. We wish to decide which alternative is best.

In order to obtain some information, the manager of Zip Company recorded the number of requests for rental cars each day over a typical period of 100 days. This information is shown in Table 9-1.

We can use the frequency data below as a probability model or representation of the uncertainty facing the Zip Company. That is, we can use a relative frequency in Table 9-1 as an estimate of the *probability* that the specified number of rental cars will be requested on a given day. This implies that the probability is zero for 9 or fewer rental

Table 9-1
REQUESTS FOR RENTAL CARS—ZIP CAR RENTAL COMPANY
SUMMARY FOR 100 DAYS

Number of Rental Cars Requested	Frequency: Number of Days	Relative Frequency
9 or fewer	0	0
10	5	0.05
11	5	0.05
12	10	0.10
13	15	0.15
14	20	0.20
15	25	0.25
16	15	0.15
17	5	0.05
18 or more	0	0
	100	1.00

requests; the probability is 0.05 for exactly 10 rental requests; etc. Note that we are restricting the possible events to between 10 and 17 rentals requested.

The use of these frequencies as a probability distribution implies a sort of "betting" model of reality. That is, we can conceive of a roulette wheel with 100 possible slots. Five of these slots are labeled "10"; five are labeled "11"; ten are labeled "12"; etc., corresponding to the frequencies, or estimated probabilities, in Table 9-1. Hence, the event "10" has only 5 chances in 100, or 1 chance in 20, of occurring, and so on. The use of these probabilities implies such a "betting" distribution about the real world.

To use the above probability distribution as a model of reality involves, of course, certain assumptions. We assume that the 100 days are a "representative" sample of past requests (i.e., there was no bias in the

manner in which the sample was selected). We assume that the future will be the same as the past insofar as rental requests are concerned. We assume that the number of requests are independent from day to day and week to week. If these assumptions are valid, our model has some validity as a representation of the real-world situation.

Decisions Based upon Probabilities Only

When presented with the data in Table 9-1, you might be tempted to make the decision of how many cars to lease with this information alone. Some such decisions and rationalizations might be as follows:

- a. Lease 10 cars. This would guarantee that all cars leased would be rented.
- b. Lease 17 cars. This would guarantee that no rental customer would be turned away.
- c. Lease 15 cars. This is the number most frequently requested (i.e., the mode).
- d. Lease 14 cars. This is the mean or expected number requested, as shown in Table 9-2.

The objection to all of the criteria (*a* to *d*) is that they make no use of the economic information available to the decision-maker. To see why the decision must depend upon the *costs* of leasing a car and the rental price, consider the following illustrations:

1. If the cost of leasing a car were zero, then the *b* criterion above (lease 17 cars) would yield the most profitable decision.

Table 9-2

CALCULATION OF EXPECTED NUMBER OF REQUESTS
REQUESTS FOR RENTAL CARS—ZIP CAR RENTALS

X Number Requested	$P(X)$ Probability	$X \cdot P(X)$
10	0.05	0.50
11	0.05	0.55
12	0.10	1.20
13	0.15	1.95
14	0.20	2.80
15	0.25	3.75
16	0.15	2.40
17	0.05	0.85
	1.00	14.00

$$E(X) = \sum X \cdot P(X) = 14.00$$

2. If the cost of leasing a car were equal to the rental price, then the *a* criterion (or the alternative of going out of business) would be the least costly alternative. It would involve zero profit, which would be preferable to the other alternatives, since they would involve losses.

From these illustrations, it appears that the economic factors such as prices and costs very much influence the correct (or most profitable) decision.

Decisions Based upon Economic Factors Only

It is possible to go to the other extreme and rely entirely upon economic factors, thereby ignoring the probability information. Let us consider this approach.

First, we arrange in a table the economic consequence for each event and for each possible action. Such a table is called a payout or *payoff table*. In construction of payoff tables, it is important to include only costs or profits which result from the actions and events under consideration. Thus, only "out-of-pocket" costs and revenues are relevant. Overhead charges and depreciation should be excluded, since they do not represent actual flows of funds. Table 9-3 is a payoff table for this problem.

Table 9-3
PAYOFF TABLE
PROFITS (IN DOLLARS) FOR ZIP CAR RENTALS

Events: Number of Rental Cars Requested	Actions: Number of Cars Leased							
	10	11	12	13	14	15	16	17
10	30	23	16	9	2	-5	-12	-19
11	30	33	26	19	12	5	-2	-9
12	30	33	36	29	22	15	8	1
13	30	33	36	39	32	25	18	11
14	30	33	36	39	42	35	28	21
15	30	33	36	39	42	45	38	31
16	30	33	36	39	42	45	48	41
17	30	33	36	39	42	45	48	51

Recall that Zip Company leased cars for \$7 per day and rented them in turn for \$10 per day. From this we can derive the profit (or loss) in the table for each combination of action and event. Thus, if Zip Company leased 13 cars and rented 11 to customers, the profit would be

$11 \times \$10$ (i.e., $\$110 = \text{revenue}$) $- 13 \times \$7$ (i.e., $\$91 = \text{cost}$), or $\$19$. We assume that there is no penalty cost (except for lost profit) when a customer requests a rental car and one is not available. The customer can be served by a competing rental agency.

Table 9-3 shows that the actions the Zip Company can take vary somewhat as to risk. The action "lease 10 cars" guarantees a profit of $\$30$ regardless of what happens. In this sense, it is the least risky or most conservative action available.² In contrast, the action "lease 17 cars" is the most risky alternative in the sense that the possible profits range from a loss of $\$19$ (when only 10 cars are rented) to a profit of $\$51$ (when all 17 cars are rented).

Most decision-makers would balk at the prospect of making a decision with only the information shown in Table 9-3. They would insist on knowing something about how "likely" the occurrence was of each possible event. The alternative "lease 10 cars" would generally be preferred if there were only a slight chance (say one in 100) that more than 10 rentals would be requested. Similarly, the alternative "lease 17 cars" would generally be preferred if requests were only rarely fewer than 17 rentals.

A person's preference or aversion to risky alternatives may depend upon how much he subjectively values the dollar amounts shown in Table 9-3. If a loss of $\$10$ or more may cut his working capital seriously, the decision-maker would avoid the alternatives "lease 16 cars" and "lease 17 cars," even though it might be very unlikely that the number of rental requests could be as low as 10 or 11. On the other hand, if profits of at least $\$40$ were needed to satisfy a certain goal (e.g., to pay off a pressing debt), the decision-maker might consider leasing upward of 13 cars only. Factors that affect the subjective worth of a gain (or loss) of a certain amount of money do influence the decision process. We shall consider such effects in detail in a later section. For now, the assumption is that no factors would subjectively change the value of money to the decision-maker; that is, a gain of $\$20$ is worth twice as much to the decision-maker as a gain of $\$10$.

Expected Monetary Value as a Decision Criterion

Both the probability information and the economic information are necessary for rational decision-making under uncertainty. The proce-

² The choice of the alternative with the highest minimal profit level is called a maximin strategy (maximizing the minimum profit). If the table is expressed in losses (negative profits), then the criterion is called minimax (i.e., select the alternative with the least [minimum] maximum loss). See references to Luce and Raiffa, Chernoff and Moses, and others on page 247 for a discussion of these types of decision strategies.

cedure for incorporating both sets of information is the subject of this section. We begin by computing the *expected monetary value* for each alternative decision. Table 9-4 illustrates this computation for the action "lease 15 cars."

The column labeled "Profit" in Table 9-4 is the profit that would result for various numbers of rental requests if 15 cars were leased (see Table 9-3). The maximum profit is \$45 when all 15 cars (or more) are requested for rental. If only 10 rentals are requested, there will be a loss (negative profit) of \$5.

Table 9-4
CALCULATION OF EXPECTED MONETARY VALUE
FOR ACTION: LEASE 15 CARS

Event: No. of Rental Cars Requested (X)	Probability $P(X)$	Profit π	Expected Profit $\pi \cdot P(X)$
10	0.05	-\$ 5	-\$ 0.25
11	0.05	5	0.25
12	0.10	15	1.50
13	0.15	25	3.75
14	0.20	35	7.00
15	0.25	45	11.25
16	0.15	45	6.75
17	0.05	45	2.25
	1.00		\$32.50

$$\text{Expected Profit} = \text{EMV} = \sum \pi \cdot P(X) = \$32.50$$

The expected monetary value (abbreviated EMV) or expected profit is interpreted in the same manner as the expected value of a random variable, $E(X)$. It is the average profit that would result if this decision were repeated many times, and each time the decision-maker chose the same alternative (in this case, "lease 15 cars"). It is the profit that is to be "expected" in the long run even though the decision is to be made only once. It is simply a weighted average profit, the weights being the probabilities of the various events. Note that a profit of \$32.50 can never occur on any day, even though the EMV is \$32.50. The actual profit that will result will be one of the values in the "Profit" column of Table 9-4.

The expected monetary value for each alternative can be computed by the procedure illustrated in Table 9-4. These values are shown in Table 9-5. The alternative "lease 13 cars" has the highest EMV. *Our*

criterion for decision-making under uncertainty is to pick that action with the highest expected profit (i.e., highest EMV).³

A little reflection should convince even the skeptical reader that this criterion is reasonable. If the decision were to be repeated day after day, the action "lease 13 cars" would bring the highest average profit. Even if the decision were a "one-shot" affair, the action "lease 13 cars" would be the "best bet" that could be made. Recall that the use of probabilities as a model of the real world implied a betting distribution for the decision-maker, the odds on various events occurring being represented by the probabilities. The action which maximizes the expected value is simply the most sensible bet or gamble in the face of the stipulated odds or probabilities.

Table 9-5

EXPECTED MONETARY VALUE (EXPECTED PROFIT)
FOR ALL ALTERNATIVES

Action: Number of Cars Leased	Expected Monetary Value (Expected Profit)
10	\$30.00
11	32.50
12	34.50
13	35.50
14	35.00
15	32.50
16	27.50
17	21.00

Note that the decision selected (lease 13 cars) is not the one suggested by any of the criteria using the probabilities by themselves or using the economic information alone. The number of cars to lease is neither the mean (which is 14) nor the mode (which is 15).

Oil-Drilling Example. An oil company is about to drill 10 wells in an isolated part of the Middle East. A certain piece of equipment is used on each well and is subject to accidental breakage. The question arises as to how many (if any) spare parts the company should transport to the drilling site.

This particular part costs \$50. If the parts are shipped with the original expedition, they will cost an additional \$50 each to ship, or a

³ Later, we shall discuss maximization of expected utility, where utility is a measure of risk evaluation. For the present, we are assuming a linear utility function for money (i.e., no aversion or preference for risk).

Table 9-6

PAYOFF TABLE FOR DECISION ON SPARE PARTS
(HUNDREDS OF DOLLARS COST)

Event: No. of Spare Needed	Actions: No. of Spares Initially Transported								
	0	1	2	3	4	5	6	7	8
0	0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0
1	5.5	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0
2	11.0	6.5	2.0	3.0	4.0	5.0	6.0	7.0	8.0
3	16.5	12.0	7.5	3.0	4.0	5.0	6.0	7.0	8.0
4	22.0	17.5	13.0	8.5	4.0	5.0	6.0	7.0	8.0
5	27.5	23.0	18.5	14.0	9.5	5.0	6.0	7.0	8.0
6	33.0	28.5	24.0	19.5	15.0	10.5	6.0	7.0	8.0
7	38.5	34.0	29.5	25.0	20.5	16.0	11.5	7.0	8.0
8	44.0	39.5	35.0	30.5	26.0	21.5	17.0	12.5	8.0
9	49.5	45.0	40.5	36.0	31.5	27.0	22.5	18.0	13.5
10 or more*	55.0	50.5	46.0	41.5	37.0	32.5	28.0	23.5	19.0

* The costs are for 10 spares needed. This is an adequate approximation since the Poisson probability of more than 10 is less than 0.0005. See Table 9-7.

total of \$100. If parts are needed later, they will have to be shipped by air at a cost of \$500 each to transport, for a total of \$550, including the cost of the part itself. At the end of the drilling operation, all parts are to be abandoned.

From the above economic information, we can draw up Table 9-6 as a payoff table. We shall restrict our alternative actions to carrying from zero to 8 spares.

Table 9-7

POISSON PROBABILITY DISTRIBUTION FOR $m = 3.0$

Event: Number of Breakages X	Probability $P(X)$
0	0.050
1	0.149
2	0.224
3	0.224
4	0.168
5	0.101
6	0.050
7	0.022
8	0.008
9	0.003
10 or more	0.001
	1.000

Table 9-8
CALCULATION OF EXPECTED COST
ACTION: CARRY 4 SPARES

Event: No. of Breakages X	Probability $P(X)$	Cost c	$c \cdot P(X)$
0	0.050	\$ 400	\$ 20.00
1	0.149	400	59.60
2	0.224	400	89.60
3	0.224	400	89.60
4	0.168	400	67.20
5	0.101	950	95.95
6	0.050	1,500	75.00
7	0.022	2,050	45.10
8	0.008	2,600	20.80
9	0.003	3,150	9.45
10 or more	0.001	3,700	3.70
	1.000		\$576.00

$$\text{Expected Cost} = \sum c \cdot P(X) = \$576.00$$

The values in Table 9-6 are the costs of purchasing and transporting the indicated number of spare parts.

Let us suppose that the drilling company knew from past experience that, on the average, 0.30 parts broke per well drilled. This is the expected breakage. Further, breakage was generally accidental (i.e., random) and did not depend upon how long a part had been in service. Since there were 10 wells to be drilled, the expected breakage would be 3 parts (0.30×10). The conditions specified above satisfy the assumptions of the Poisson process. Thus, we could use the Poisson distribution as our model or representation of the real world (i.e., as our betting distribution about what event will occur). The Poisson distribution with $m = 3.0$ is shown in Table 9-7 (taken from Appendix H).

A sample calculation for the expected cost of the action "transport 4 spares" is shown in Table 9-8.

The expected cost for each action is shown in Table 9-9. The action "carry 5 spares" has the minimum expected cost. (In this example, minimization of cost is equivalent to maximization of profit.) Hence, this is the optimal decision.

SUBJECTIVE PROBABILITIES AND DECISION-MAKING

In the two above examples, we were able to build without much difficulty a probability model of the world. In the first example, we had

Table 9-9
EXPECTED COST FOR EACH ALTERNATIVE

Action: Number of Spares	Expected Cost
0	\$1,650.55
1	1,230.75
2	887.50
3	670.15
4	576.00
5	574.25
6	628.05
7	709.35
8	802.75

available historical frequency data which served adequately as our probability distribution. In the second example, we found that breakage of parts satisfied the assumptions of the Poisson process, so we used a Poisson probability distribution. In many important decision situations, such ready-made probability distributions are not available. Consider the following decision situation: A manufacturer is trying to decide upon the size of plant to build for a new product. The market for the new product is quite uncertain. Small quantities to satisfy demands over the next two years can be produced with present facilities. But a new plant will be needed and must be started now to be completed in two years.

The ultimate market demand, let us say, will be either high or low, but the decision-maker does not now know which it will be.

There are two actions open to management, (1) build a large plant, which would be suited to high demand, or (2) build a small plant, which would be best suited to low demand. A large plant would cost \$4 million; a small plant, \$2 million. Profits (excluding cost of the plant) for the large plant would be \$10 million in the case of high demand and \$5 million in the case of low demand. For the small plant the

Table 9-10
NET PROFITS FROM BUILDING VARIOUS SIZE PLANTS
(IN MILLIONS OF DOLLARS)

Event:	Action:	
	Build Large Plant	Build Small Plant
High Demand	$10 - 4 = 6$	$6 - 2 = 4$
Low Demand	$5 - 4 = 1$	$5 - 2 = 3$

profits (excluding plant cost) would be \$5 million for low demand but only \$6 million for high demand due to capacity limitation of the small plant. These figures are summarized in Table 9-10.

The manufacturer could gather opinions of various experts before making his decision. Market research specialists could be consulted, as well as economists. The latter could predict future economic levels, upon which market demand would depend. But often there is disagreement, even among experts about future market demand or about the general level of the whole economy.

He could pick his favorite forecaster and base his decision upon this educated opinion. Better yet, he could estimate probabilities or "betting odds" that the events—high or low demand—will occur. These probabilities would be based upon the decision-maker's judgment, taking into account all available information. For example, the executive may assign the following probabilities.

<i>Event</i>	<i>Probability</i>	<i>Odds</i>
High Demand	0.60	3 to 2 (i.e., 3 chances in 5)
Low Demand	0.40	2 to 3 (i.e., 2 chances in 5)

Such probabilities imply that he should be willing to bet on high demand at odds of 3 to 2 and on low demand at odds of 2 to 3.

With these probabilities, the expected profit of building the large plant is $(0.60)(6) + (0.40)(1) = \4.0 millions. The expected profit of building the small plant is $(0.60)(4) + (0.40)(3) = \3.6 million. The action "build the large plant" is preferred at the indicated odds. In fact, as long as the executive assigns a probability greater than $\frac{1}{2}$ to the event "high demand," building the large plant has the higher expected profit. (If the probability of high demand and of low demand are each $\frac{1}{2}$, the expected profit of the two alternatives is the same—\$3.5 million.) Thus, building the large plant would be preferred as long as the manufacturer felt that high demand was more likely than low demand. As in our previous examples, it is a combination of probabilities and economic information that is relevant to decision-making under uncertainty.

In this example, the probabilities merely reflect the uncertainty in the mind of the decision-maker. Hence, they are referred to as *subjective probabilities*. Different decision-makers would likely have different sets of probabilities about the same events. We do not propose, in this text, to dwell upon how subjective probability distributions should be formulated, other than the common sense advice of carefully examining past information and expert opinion. Rather, we hope to show how to

act (i.e., make decisions) in a manner consistent with one's subjective probabilities.

In a sense, all probability models of the real world are subjective. The decision-maker must have confidence that the model adequately represents the world or he cannot use the model as a basis for his decisions. In the example of the Zip Car Rental Company above, the decision-maker must assume that the future probabilities can be represented by historical frequencies. This is a subjective assumption. The decision-maker could have assumed just as easily that the historical frequencies needed to be modified. A modified version of Table 9-1 could then show the appropriate subjective probabilities.

DECISION TREES

In the example in the previous section, the manufacturer had only a *single* decision to make—he could build either a large or a small plant. Subsequent market conditions would determine what profit he would make.

Suppose it is possible for the manufacturer to build a small plant and expand it at a later date when the market demand for the new product is known. The cost of such an expansion would be \$3 million. The expanded facilities could enable the firm to meet the sales requirements for a high level market demand and hence to obtain the same \$10 million profits (excluding plant cost) that could be obtained by a large factory.

Note that in this revised example, the manufacturer is making a *sequence* of decisions: first the decision—large versus small plant; and second, at a later date, the decision—expand or not expand the small plant (if he chose the small plant for the first decision). In between these decisions, the manufacturer obtains new information; that is, he discovers whether the market demand will be high or low. The manufacturer may improve his first decision, therefore, by taking account of the possibilities offered in the second decision.

Sequential Decisions and Decision Trees

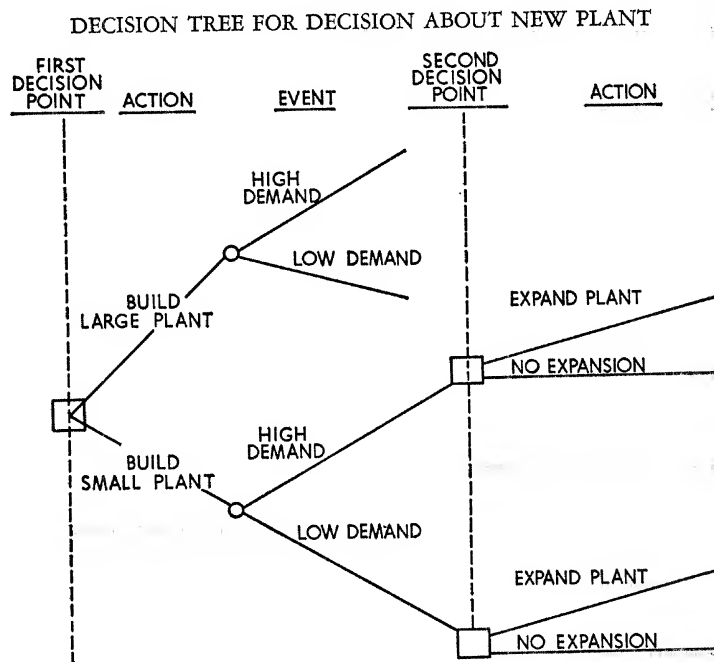
One method of analyzing problems which involve a sequence of decisions is to express the alternatives in the form of a *decision tree*. The decision tree for the problem faced by the manufacturer is shown in Chart 9-1.

Starting at the left, the first two lines or *branches* of the decision tree represent the decision alternatives for the first decision—either build a large or a small plant. At the end of each of the decision (or action)

branches comes a fork with two branches representing the events high and low market demand for the new product. It is unknown at the time the first decision (size of plant) must be made which of these event branches will actually occur.

For the "build large plant" action, the tree ends after the event branches. However, for the "build small plant" action a second decision point is reached *after* each of the events "high demand" or "low demand." The decision-maker can choose between the actions "expand the

Chart 9-1



plant" and "no expansion" after he knows the market demand level. These actions are represented as branches on the decision tree. Including both action branches after each of the forks at the second decision point may seem unnecessary at first. One would generally expect to expand the plant in response to high demand and not to expand if low demand materialized. But we cannot be sure of this until we include the economic information in the tree, which we shall do below. There always is the possibility, for example, that the expansion will cost more than the additional revenue even from high market demand. Hence, we should retain both action alternatives at each of the second decision points.

The decision tree as shown in Chart 9-1 represents the basic structure of this decision problem. The decision actions and the uncertain or chance events are shown; and the order in which various actions precede or follow events is indicated.

Analysis Using Decision Trees

Once we have set up a decision problem in the form of a tree, the next step is to analyze the problem and arrive at a solution.

Economic Information and Probabilities. The costs or profits of various actions and the likelihoods or probabilities of various events must be incorporated in the analysis just as was done with payoff tables in the earlier parts of this chapter. The probabilities for various events can be shown alongside each event branch as is illustrated in Chart 9-2, where the probabilities are 0.6 that high demand will materialize and 0.4 for the low-demand possibility.

The economic consequences or payoffs are also determined as before. They represent the net cash outflow or inflow for various action-event combinations. In Chart 9-2, the payoffs are represented at the end of final branch of the tree. For a large plant and high demand, the net cash inflow is \$6 million; and if demand is low, the payoff is \$1 million. These are exactly the figures shown in Table 9-10. Similarly, if a small plant is built initially and no expansion is made, the amounts \$4 million and \$3 million shown in Chart 9-2 are again the figures in Table 9-10 for high and low demand, respectively. The payoff or net profit of \$5 million related to expanding the plant with high demand is determined as follows:

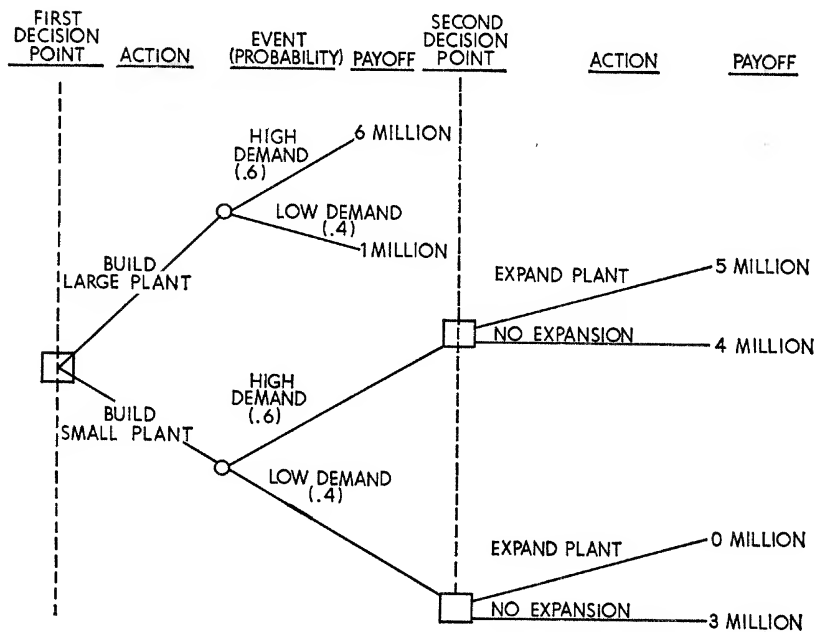
Profit from high demand (with production ability to meet demand)		\$10 million
Less: Cost of building small plant	\$2 million	
Cost of expanding	<u>3 million</u>	
Total cost		<u>5 million</u>
Payoff		\$ 5 million

Similarly, expanding in the face of low demand costs the \$5 million as above and only gives \$5 million in profit for a net payout of 0, as shown at the end of the "Small Plant—Low Demand—Expand" branch in Chart 9-2.

Working Backward on the Decision Tree. With the payoffs and probabilities shown on the decision tree, the next step is to begin the analysis with the aim of finding that decision (or sequence of decisions) which is best. To do this, *we begin by working backward on the tree,*

Chart 9-2

DECISION TREE FOR DECISION ABOUT NEW PLANT
(INCLUDING PROBABILITIES AND PAYOFFS)



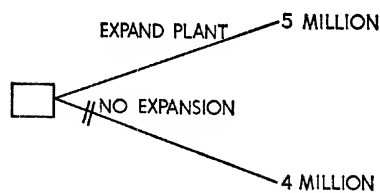
from the final or end branches back toward the first decision point.

The second decision point is thus the first considered. At the end of the high demand branch is the fork shown in Chart 9-3, Panel A. Since the action "Expand the Plant" leads to \$5 million net profit as opposed to only \$4 million for no expansion, that alternative is selected. The "no expansion" branch is removed from further consideration by drawing two lines through it, as shown. Similarly, for the decision at the end of the low demand branch, Chart 9-3, Panel B, the action "no expansion" is preferred (with net profit \$3 million), and the action

Chart 9-3

DECISIONS AT END BRANCHES

PANEL A



PANEL B

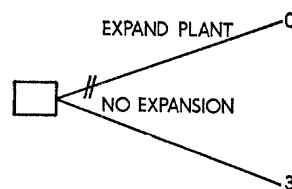
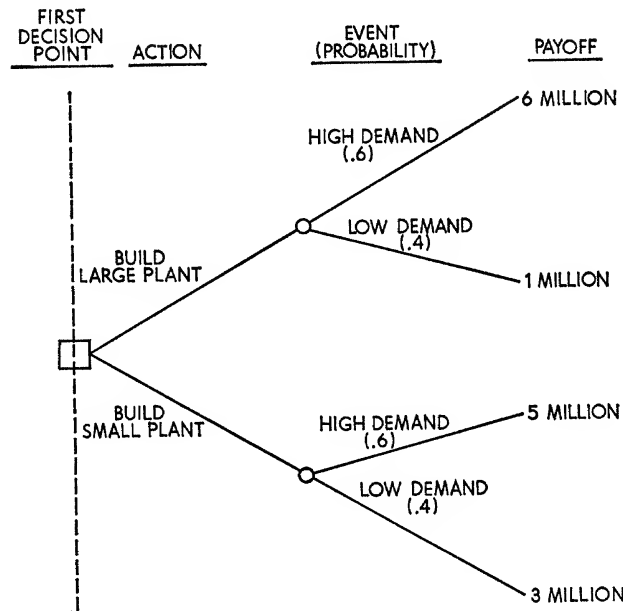


Chart 9-4

REDUCED DECISION TREE

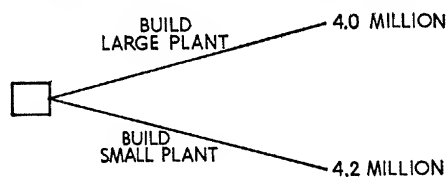


"expand plant" is eliminated. The reduced decision tree appears in Chart 9-4. This completes the analysis for the second decision point.

We now move backward to the "event" forks, with branches labeled "high demand" and "low demand," respectively. At each of these forks an expected value is taken using the payoffs at the ends of the branches and the probabilities shown. For the fork at the end of the "Build Large Plant" action the expected value is \$4.0 million ($\$6 \text{ million} \times 0.6 + \$1 \text{ million} \times 0.4$), the same as obtained from the payoff table analysis of Table 9-10. For the fork at the end of the "Build Small Plant" branch, the expected value is \$4.2 million ($\$5 \text{ million} \times 0.6 + \$3 \text{ million} \times 0.4$). By replacing the event forks by their expected values, the final reduced form of the decision tree is obtained (Chart 9-5).

Chart 9-5

FINAL REDUCED DECISION TREE



The best decision for the manufacturer, therefore, is to build the small plant now and to decide upon expansion later when market demand is known.

Discussion. The only immediate decision facing the manufacturer was the one involving the initial size of the plant. But in order to make this decision, he had to take account of the possibility of a subsequent decision on expansion. Thus, he makes a *sequence* of two decisions—(1) build a small plant and (2) expand if a large market potential materializes—rather than a *single* decision. Compare the result of this analysis with the decision to build a large plant (Table 9-10) when only the single decision was considered.⁴ This earlier conclusion was just opposite to the sequential decision to build, initially, a small plant.

A Further Example

To illustrate the use of the decision tree in a more complex situation, consider the following example: Artex Computers is interested in developing a new tape drive for a proposed new computer. Artex does not have research personnel available to develop the new drive itself and so is going to subcontract the development to an independent research firm. Artex has set a fee of \$250,000 for developing the new tape drive and has asked for bids from various research firms. The bid is to be awarded not on the basis of price (set at \$250,000) but on the basis of both the technical plan shown in the bid and the reputed technical competence of the firm submitting the bid.

Boro Research Institute is considering submitting a proposal (i.e., a bid) to Artex Computer to develop the new tape drive. Boro Research management estimated that it would cost about \$50,000 to prepare a proposal; further, they estimated that the chances were about 50-50 that they would be awarded the contract.

There was a major concern among Boro Research engineers concerning exactly how they would develop the tape drive if they were awarded the contract. There were three alternative approaches that could be tried. One approach involved the use of certain electronic components. The engineers estimated that it would cost only \$50,000 to develop a prototype (i.e., a test version) of the tape drive using the electronic approach, but that there was only a 50 percent chance that the proto-

⁴ It would have been possible to modify the original payoff table to include the possibilities of an expanded small plant. Indeed, we can always reduce decision trees to appropriate payoff tables by careful definition of actions and events. However, it is generally easier to analyze a sequential problem by using a decision tree rather than a single payoff table.

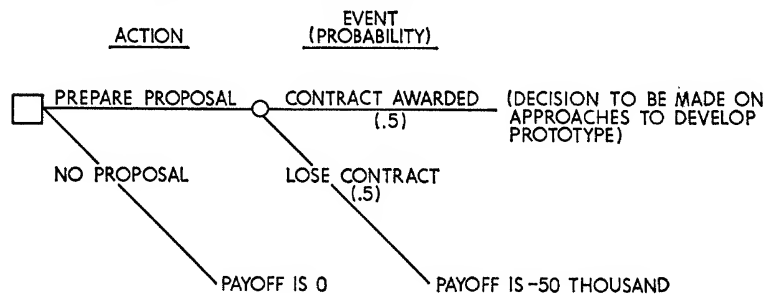
type would be satisfactory. A second approach involved the use of certain magnetic apparatus. The cost of developing a prototype using this approach would cost \$80,000 with 70 percent chance of success. Finally, there was a mechanical approach with cost of \$120,000, but the engineers were certain they could develop a successful prototype with this approach.

Boro Research could have sufficient time to try only two approaches. Thus, if either the magnetic or electronic approach were tried and failed, the second attempt would have to use the mechanical approach in order to guarantee a successful prototype.

The management of Boro Research was uncertain how to take all this information into account in making the immediate decision—

Chart 9-6

BORO RESEARCH INSTITUTE
DECISION ON PREPARATION OF PROPOSAL



whether to spend \$50,000 to develop a proposal to send to Artex Computers.

Since this decision problem seems complex, let us build the decision tree in steps. The first decision facing Boro Research involves the actions "Prepare a Proposal" and "Do Not Prepare a Proposal." If a proposal is developed and submitted to Artex Computers, then either of the events "Contract Awarded to Boro Research" or "Boro Research Loses Contract" must occur. Each event has the probability 0.5. These choices are shown in Chart 9-6.

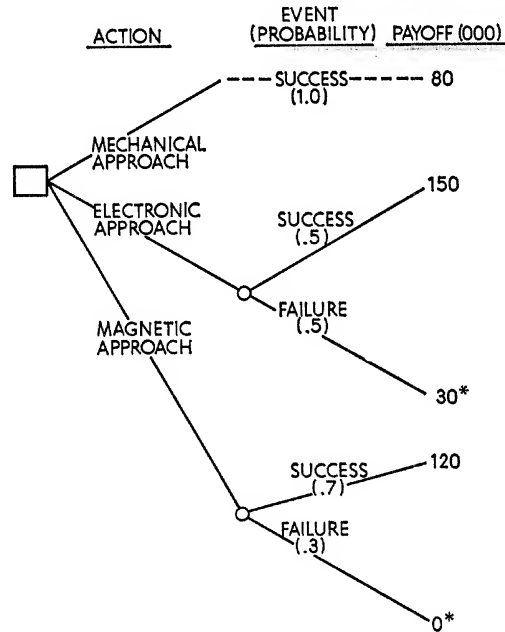
If Boro Research decides *not* to prepare a bid, the net payoff is zero. If a bid is prepared but the contract is lost, Boro Research loses the \$50,000 cost of preparing the bid (i.e., the payoff is $-\$50,000$). If the contract is awarded to Boro Research, then the next decision—the choice between alternative methods of developing a successful tape drive—must be made.

In the second decision, Boro Research must decide which of the three approaches—mechanical, electronic, or magnetic—to try first.⁵ This decision is shown in Chart 9-7.

If the mechanical approach is selected, a successful prototype will be developed for sure and Boro Research will have a net return of \$80,000 (\$250,000 value of contract minus \$50,000 proposal cost minus \$120,000 to develop the mechanical prototype). If either of the other approaches is selected, it may succeed or fail. Failure means that the

Chart 9-7

BORO RESEARCH INSTITUTE
DECISION ON WHICH APPROACH TO TRY FIRST



* Mechanical approach must be used.

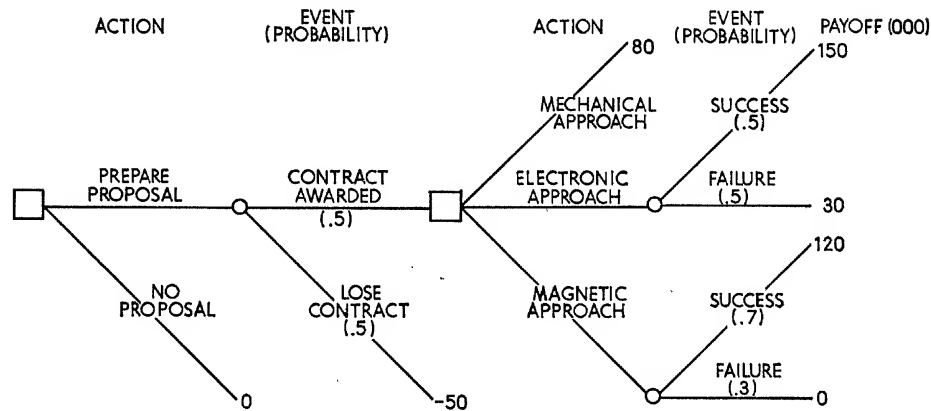
mechanical approach must be used in order to guarantee a successful prototype within the time available. The payoffs shown in Chart 9-7 are calculated as follows:

⁵ Boro Research could possibly add a fourth alternative—develop both the electronic and magnetic prototypes simultaneously and follow with the mechanical only if both fail. This could be added as a branch of the tree. However, the cost of this would be at least \$180,000 (more if neither approach produced a success), and this is greater than the cost of a mechanical prototype (\$170,000).

		Payoff (Thousands of Dollars)					
		<div> <div>Cost of</div> <div>Proto-</div> <div>type</div> <div>Indi-</div> <div>cated</div> </div> <div> <div>Cost of</div> <div>Proto-</div> <div>type</div> <div>Mechani-</div> <div>cal Proto-</div> <div>type</div> </div>					
<i>End of Branch</i>	<i>Fee</i>						
Electronic Approach							
Success.....	250	—	50	—	50		= 150
Failure.....	250	—	50	—	50	— 120	= 30
Magnetic Approach							
Success.....	250	—	50	—	80		= 120
Failure.....	250	—	50	—	80	— 120	= 0

Chart 9-8

COMPLETE DECISION TREE FOR BORO RESEARCH INSTITUTE



The complete decision tree is shown as Chart 9-8. It is obtained by joining Charts 9-6 and 9-7.

Working Backward. The expected values are calculated for each of the event forks in the far right part of the tree. Thus, the expected payoff associated with the electronic approach is \$90,000 (0.5×150 plus $0.5 \times 30 = 90$) and for the magnetic approach is \$84,000 (0.7×120 plus $0.3 \times 0 = 84$). These expected payoffs are inserted in circles beside the appropriate forks in Chart 9-9.

Moving left to the decision point, we see that the electronic approach offers the highest expected payoff (\$90,000) and is the best choice. The value \$90,000 is written (circled) beside the decision point and the nonpreferred approaches are indicated by drawing || on the branches.

The tree now has a payoff of +\$90,000 if the contract is awarded

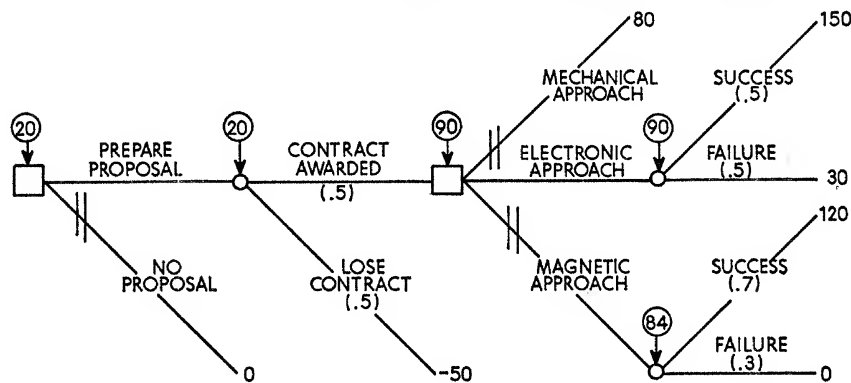
and $-\$50,000$ if not. The expected value of preparing a proposal is $\$20,000$ (0.5×90 plus $0.5 \times (-50) = 20$). This is written in a circle beside the event fork.

Finally, the choice must be made between the expected payoff of $\$20,000$ for preparing the proposal and zero if the proposal is not prepared. The first, of course, is selected, and the mark $||$ drawn through the "No Proposal" branch.

In summary, Boro Research should prepare the proposal, anticipating $\$20,000$ as the expected value of this decision. If the contract is awarded, the electronic approach should be tried first; but if this fails, the mechanical approach must be used.

Chart 9-9

BORO RESEARCH INSTITUTE ANALYSIS OF DECISION TREE



RISK IN DECISION-MAKING: THE UTILITY OF MONEY

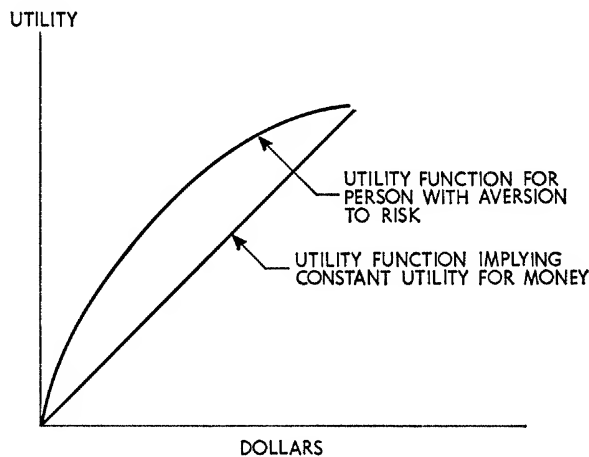
Expected *monetary* value is not always the best criterion to use in decision-making. If you were offered your choice of one of two alternatives: either (a) a 50-50 chance of $\$250$ or zero or (b) $\$100$ for sure, you would probably take the $\$100$. Most people would, despite the fact that the expected monetary value of the 50-50 gamble is $\$125$. Is this evidence in conflict with the decision criterion which we expressed in an earlier section—the criterion that one should pick the decision alternative with the highest expected monetary value? Yes, it is! And we are now in a position to extend or elaborate upon our measure of value. The problem arises because the value of money to people is not always a linear function of the amount of money. Generally, $\$200$ is not worth twice as much to a person of modest means as $\$100$. It would matter a great deal to you whether I gave you zero or $\$100$; but it probably

would not matter a great deal if the choice were between \$1,000,000 and \$1,000,100. This is because money has *diminishing utility* to most of us; the first \$100 we receive is most important, while successive increments of \$100 have less and less subjective value.

We see the same phenomenon at work when people buy insurance. For most people, insurance is bound to be a "bad bet" from a purely monetary point of view, since the insurance company must pay its expenses and make a profit in addition to covering the risk. That is, the expected monetary value of insurance is negative, from the buyer's viewpoint. However, most of us are willing to pay a small amount (the

Chart 9-10

TYPICAL UTILITY FUNCTIONS



insurance premium) to guard against a disastrous occurrence, even though the chance of such an event happening may be quite small.

In order to make decisions under uncertainty, we must have some way to measure a decision-maker's attitude toward risk and express this in quantitative terms. The appendix at the end of this chapter gives a brief discussion of how this can be done. The result is a function relating dollar amounts to a measure of *utility*.⁶ A typical function is shown in Chart 9-10.

For a person who has an aversion to risk (e.g., one who would prefer \$100 for sure to a 50-50 chance at zero or \$250), the shape of the function would reflect his diminishing utility of money, as shown. A

⁶ The word "utility" is somewhat misleading. It is merely a risk equivalence measure and bears no direct relationship to "utility" as commonly used in economic theory. The utility scale (the ordinate in Chart 9-10) is not unique. (The scale can be multiplied by a constant or shifted up or down without changing the function in any real sense.)

person who is willing to use expected monetary value would have a linear utility function. (He'd be indifferent as between the alternatives of a certain \$125 and a 50-50 chance of zero or \$250.)

In many decision situations, the amounts of money involved are small relative to the resources of the decision-maker. Thus, for inventory decisions that involve only a few thousand dollars, a large corporation would use expected monetary value. Over this range (plus or minus a few thousand dollars), the utility function for the company is approximately linear. For more important decisions (e.g., the decision to build a new factory or to enter a new market), monetary value alone is generally not appropriate. In such situations, the decision-maker should determine his utility for money (as shown in the Appendix at the end of this chapter). The decision criterion is then to pick the alternative with the highest expected *utility*, rather than the highest expected monetary value.

SUMMARY

This chapter described a procedure for making decisions in an uncertain environment. The procedure, in skeletal form, involved:

1. Defining the possible events that can occur.
2. Defining the actions that can be taken.
3. Determining the value (in dollars or utility) of each action-event combination.
4. Describing the decision-maker's uncertainty about the events by a set of probabilities.
5. Finding the expected value of each alternative action by multiplying its value for each event by the probability and summing.
6. Selecting that alternative with the highest expected profit (or utility).

To specify this decision procedure is merely to organize the decision-making process in a systematic and logical fashion. No one making a decision under uncertainty can avoid the steps listed as 1 through 6 above—though he might do some steps in an intuitive manner. Our procedure is no more than a completely specified logical framework.

Decision trees may be used to analyze problems that involve a sequence of decisions. The various actions that may be taken are shown on the tree as branches emanating from a fork, and the various events that may occur are similarly represented. Hence, the tree diagram ties together a sequence of decisions and events.

The payoffs for various sequences of actions and events are shown at the end branches of the tree. And the probabilities for the various events are listed below each event.

The decision tree is analyzed by working backward from the final action or event to the first action to be chosen. At each stage an expected value is calculated over possible events; and a choice is made among alternative actions, selecting the one with the highest expected value.

Utility values may be substituted for monetary values, for those whose subjective value of money is not linear, by methods described in the appendix of the chapter.

In the chapters that follow, we shall extend this analysis. We shall examine, first, the possibility of postponing the decision while additional information is collected (Chapter 10). Subsequently (Chapters 15 and 16), we shall consider obtaining information by sampling.

APPENDIX: DERIVATION OF UTILITY CURVES FOR DECISION-MAKING UNDER UNCERTAINTY

Suppose a businessman had a choice of one of two contracts. The profit resulting from either contract is uncertain. The contracts and their probabilities and payoffs are:

Contract I			Contract II		
<i>Event</i>	<i>Proba- bility</i>	<i>Payoff</i>	<i>Event</i>	<i>Proba- bility</i>	<i>Payoff</i>
<i>A</i>	0.30	+\$9,000	<i>Q</i>	0.25	+\$7,500
<i>B</i>	0.45	+ 6,000	<i>R</i>	0.60	+ 2,000
<i>C</i>	0.25	- 9,000	<i>S</i>	0.15	- 5,000
EMV = +\$3,150			EMV = +\$2,325		

It is easy enough to calculate the expected monetary value of each contract shown above. In order to decide which contract the businessman prefers, however, we intend to ask him a series of questions. The questions are intended to measure his preferences in risk situations simpler than the above contracts.

We begin by selecting two reference points. One should be larger than the largest positive money value in the real decision problem. For this upper reference point, let us arbitrarily choose \$10,000. The other reference point should be less than the lowest money value in the real problem; let us select -\$10,000 for this reference point. We arbitrarily assign utility values of 1.0 and 0.0 to these reference points.⁷ That is,

⁷ The choice of scale is arbitrary. We could have chosen $u(+\$10,000) = 502.6$ and $u(-\$10,000) = -29$ if we wished. The use of a scale between 1.0 and 0.0 is convenient.

$$u(+10,000) = 1.0$$

$$u(-10,000) = 0.0$$

We then give the decision-maker a choice of the following kind: What is the maximum amount you would pay to be released from a contract that gives you a $1/2$ chance at $+\$10,000$ and a $1/2$ chance at $-\$10,000$?⁸

The answer to such a question would be a personal matter, depending upon the resources and the propensity for risk of the decision-maker. Let us suppose that the decision-maker said that he would be willing to pay up to $\$2,000$ to be released from the gamble (i.e., from the contract giving a $1/2$ chance at $+\$10,000$ and a $1/2$ chance at $-\$10,000$). In other words, the decision-maker is indifferent between a sure amount of $-\$2,000$ and the gamble (or contract). We postulate that the utility of $-\$2,000$ is equivalent to the expected utility of the contract:

$$\begin{aligned} u(-\$2,000) &= 1/2u(+\$10,000) + 1/2u(-\$10,000) \\ &= 1/2(1.0) + 1/2(0.0) = 0.5 \end{aligned}$$

Hence, our utility index for $-\$2,000$ is 0.5. Using this figure, we can proceed to ask further questions. We might ask: What is the minimum amount the decision-maker would accept for a contract that gave him a $1/2$ chance for $+\$10,000$ and a $1/2$ chance for a $-\$2,000$?⁹ Suppose the answer is $+\$2,000$. We then determine the utility index for $+\$2,000$ as

$$\begin{aligned} u(+\$2,000) &= 1/2u(+\$10,000) + 1/2u(-\$2,000) \\ &= 1/2(1.0) + 1/2(0.5) = 0.75 \end{aligned}$$

We can continue asking similar questions:¹⁰ At what amount is the decision maker indifferent to a contract with a $1/2$ chance of $-\$2,000$ and a $1/2$ chance at $-\$10,000$? Suppose the answer is $-\$4,000$. Then,

$$\begin{aligned} u(-\$4,000) &= 1/2u(-\$10,000) + 1/2u(-\$2,000) \\ &= 1/2(0.0) + 1/2(0.5) = 0.25. \end{aligned}$$

⁸ The contract may have positive value in which case the question should be: What is the minimum amount (positive) that you would accept to sell the contract to someone else?

⁹ The question would be worded, "How much would he pay to get out of a contract . . ." if the contract had negative value (less than zero dollars).

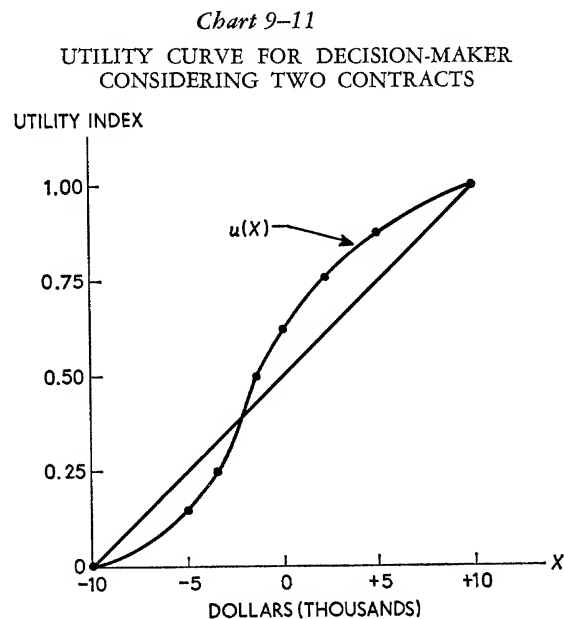
¹⁰ An alternative procedure is to hold the amounts in the question constant (i.e., keep the $+\$10,000$ and $-\$10,000$) but change the odds for each question. The utility index is determined in the same manner.

Suppose we continued and determined more answers. These are shown, together with the ones discussed above, in the table:

<i>Chance</i>	<i>Gamble</i>	<i>Indifference Amount</i>	<i>Utility Value</i>
1/2	+\$10,000	-\$2,000	$u(-\$2,000) = 0.5$
1/2	-\$10,000		
1/2	+\$10,000	+\$2,000	$u(+\$2,000) = 0.75$
1/2	-\$ 2,000		
1/2	-\$10,000	-\$4,000	$u(-\$4,000) = 0.25$
1/2	-\$ 2,000		
1/2	+\$ 2,000	-\$ 500	$u(-\$ 500) = 0.625$
1/2	-\$ 2,000		
1/2	+\$ 2,000	+\$5,000	$u(+\$5,000) = 0.875$
1/2	+\$10,000		
1/2	-\$10,000	-\$5,000	$u(-\$5,000) = 0.125$
1/2	-\$ 4,000		

The utility function is shown in Chart 9-11. A smooth curve has been drawn connecting the points determined above.

We can now return to the original situation with which we started this appendix. The two contracts are shown below, together with the corresponding utility index values. The utility values are read from Chart 9-11.



Contract I				Contract II			
Event	Proba- bility	Monetary Outcome	Utility Value	Event	Proba- bility	Monetary Outcome	Utility Value
A	0.30	+\$9,000	0.98	Q	0.25	+\$7,500	0.95
B	0.45	+ 6,000	0.90	R	0.60	+ 2,000	0.75
C	0.25	- 9,000	0.02	S	0.15	- 5,000	0.125
Expected Monetary Value = +\$3,150				Expected Monetary Value = +\$2,325			
Expected Utility = 0.704				Expected Utility = 0.706			

Contract II now has a slightly greater *utility* value, though Contract I has a much greater *monetary* value. Hence, this particular businessman should choose Contract II. Note that both contracts would be preferred to doing nothing, since $u(\$0.0) = 0.66$.

PROBLEMS

- Characterize each of the following as decision-making under certainty or uncertainty. Give your reason in one or two sentences.
 - Decision about whether or not to develop a new type of product (e.g., a new type of drug).
 - Decision about what price to put on a bid for a construction contract.
 - The price to set for a product.
 - Scheduling of production orders through a machine shop.
 - Inventory decisions.
- In each of the decision situations below, indicate in a general way what events might occur. From what sources would management obtain the probabilities of these events? To what extent are the probabilities subjective or objective?
 - The decision about the number of clerks to staff a tool crib in a factory and the effects upon the time spent by mechanics waiting for tools.
 - The marketing of a new product.
 - Company sales forecast ten years in the future.
 - The decision about the size of a new factory.
 - The decision about how many items to stock in inventory.
- Consider the following payoff table:

PAYOFF TABLE
(DOLLARS PROFIT)

Event	Probability	Actions				
		A	B	C	D	E
I	0.05	100	120	210	140	180
II	0.05	110	160	190	140	180
III	0.10	130	200	170	140	100
IV	0.30	150	180	120	140	180
V	0.40	180	150	100	140	120
VI	0.10	250	100	100	140	120

The probabilities of events I through VI are shown in the second column. Calculate the expected monetary value of each action. Which action gives the highest expected profit?

4. Suppose, in the payoff table in Problem 3, that the probabilities for events I through VI are

Event	Probability
I	0.10
II	0.40
III	0.30
IV	0.10
V	0.05
VI	0.05

Determine the expected value for each action. Which action gives the highest expected profit?

5. A merchant carries a perishable good in his inventory. The item costs \$5 each and sells for \$9. At the end of the day, any unsold items must be thrown away (no value). Assuming that demand for the item follows a Poisson distribution with mean $m = 3.0$ per day, how many items should the merchant stock on any given day? What is the expected profit?
6. Suppose in Problem 5 that the demand for the item followed this distribution:

Demand	Probability
0	0
1	0.4
2	0.3
3	0.2
4	0.1
5 or more	0
	1.0

How many items should the merchant stock? What is the expected profit?

7. A company is trying to decide what size plant to build in a certain area. Three alternatives are being considered: plants with capacities of 10,000, 15,000, and 20,000 units, respectively. Demand for the product is uncertain, but management has assigned the probabilities listed below to five levels of demand. The table below also shows the profit for each alternative and each possible level of demand for the product.

PAYOFF TABLE, SHOWING PROFITS (IN MILLIONS OF DOLLARS) FOR THE VARIOUS SIZES OF PLANTS AND LEVELS OF DEMAND

Demand in Units Z	Probability $P(Z)$	Actions: Build Plant with Capacity of:		
		10,000 Units	15,000 Units	20,000 Units
5,000	0.2	-4.0	-6.0	-8.0
10,000	0.3	+1.0	0.0	-2.0
15,000	0.2	+1.5	+6.0	+5.0
20,000	0.2	+2.0	+7.5	+11.0
25,000	0.1	+2.0	+8.0	+12.0

What size plant should be built?

8. Suppose your plant is having a new cylindrical extruder made to order by Farrell-Birmingham, a company that specializes in the manufacture of large, custom-made machinery such as this. One of the key parts in the extruder is a double-toothed pinion gear, which incurs a great deal of strain in the extruding process and is apt to break down.

Farrell-Birmingham will include extra gears, at a cost of \$2,000 a piece, when they ship the extruder to you. If, on the other hand, you do not order enough extra gears initially and have to place a new order at some later date, Farrell-Birmingham will have to prepare a new mold and will charge you a flat fee of \$14,000 for 5 extra gears.

Your plant foreman estimates that no more than 5 breakdowns of the pinion gear will occur during the life of the extruder and attaches the following probabilities to the number of failures to be expected:

No. of Breakdowns	Probability
0	0.1
1	0.2
2	0.3
3	0.2
4	0.1
5	0.1

Draw up a payoff table. How many extra gears should you order now? What is the expected cost? (Hint: remember that if you order 2 extra gears and have 3 breakdowns, you will have to place a second order.)

9. The Gusher Oil Company is considering leasing a particular parcel of land in a recently discovered oil area. The cost of the lease is \$40,000. The cost of drilling an oil well on the site is \$80,000. If oil is discovered, the net profit from the well (excluding drilling costs and the cost of the lease) will be \$360,000.

Draw up a payoff table. Assuming that Gusher maximizes expected monetary value, what is the minimal probability of finding oil necessary for Gusher to take the lease and start drilling?

10. The LMN Company produces novelty items for the Christmas season. A particular item is sold for \$1 each. Management assigns the following probabilities to various levels of sales:

Sales, Units	Probability
1,000	0.1
1,500	0.4
2,000	0.3
2,500	0.1
3,000	0.1

The cost of manufacturing this novelty item varies with the number produced as shown below:

No. Produced, Units	Average Cost per Unit, Cents
1,000	60
1,500	46 $\frac{2}{3}$
2,000	38 $\frac{3}{4}$
2,500	33 $\frac{3}{8}$
3,000	29 $\frac{1}{2}$

If more items are produced than sold, up to 1,000 units of the excess may be disposed of at a price of 10 cents each. Any additional excess items have no value. Items may be produced only in blocks of 500 units. Draw up a payoff table. How many units should be produced? What is the expected profit?

11. The credit manager of IJK Industrial Products considered extending a line of credit to Lastco Construction Company. Lastco was a new company and was definitely considered a credit risk. Based upon IJK's experience, approximately 30 percent of firms like Lastco failed within a year with a severe loss to creditors. Another 25 percent had serious financial troubles. Of the remaining 45 percent, 25 percent became sporadic customers and only 20 percent became good customers over a period of time.

Those customers that failed completely averaged sales of \$1,500 each before failing and left an average unpaid balance of \$800 which was totally lost.

Those that had severe financial troubles usually lost their credit, but only after they had made purchases of \$2,000 and had unpaid balances of \$1,000 of which half (\$500) was ultimately collected.

Firms that were sporadic customers averaged sales of only \$500 (with no credit losses). The good customers, however, averaged sales of approximately \$6,000.

IJK was concerned about granting credit to Lastco. On the one hand, if credit was not extended to a potential customer, his business was lost. On the other hand there were substantial risks of nonpayment (as described above), and since IJK made an average contribution (price minus variable cost) of only 20 percent of sales, this exaggerated the problem. In addition, there were collection costs of \$100 per customer for those that failed or were in financial trouble.

Draw up a payoff table for this decision problem. Should IJK grant credit to Lastco?

12. Suppose, for the example of Boro Research Institute described in the text (page 207), that Boro Research was not under a time constraint to produce the prototype. In this case, the firm could possibly try both the uncertain approaches (electronic and magnetic) before using the certain mechanical approach.

Draw up the decision tree for this case. How should Boro Research proceed to develop the prototype?

13. In which of the decision situations do you think maximization of expected monetary value (as opposed to expected utility) is a satisfactory decision criterion?

- a) Decision about building a new factory.
 - b) Decision about entering a new market.
 - c) Decision about buying out another company.
 - d) Decisions about production schedules.
 - e) Decisions about warehouse location.
 - f) Decisions about what quantities to order for inventory.
-

14. The Pearson Company is considering the purchase of a new machine which will be used exclusively in the production of a certain product. There are two machines on the market which would be satisfactory. Machine *A* has a purchase cost of \$10,000 and will save \$1.00 per item over the manufacturing process now used. Machine *B*, on the other hand, will cost \$60,000 but will effect cost savings of \$3.00 per item over the current cost. Both machines have a life of 5 years.

The future market is somewhat uncertain. Management expressed the following probabilities for *total* sales over the 5-year period.

Total 5-Year Sales (Units)	Probability
10,000	0.1
20,000	0.3
30,000	0.4
40,000	0.2

Ignore all discounting in your calculations. Which machine should Pearson purchase? What is the expected savings of each action?

15. The Lockjaw Company is about to bid on a contract to manufacture a large electric generator for a municipal utility company. Lockjaw has two competitors, *A* and *B*, who will be submitting competitive bids. The lowest bidder will win. If two or more bid the same lowest price, the winner will be determined by random draw.

In order to obtain some feel for how Lockjaw had fared against its competitors in the past, the company statistician prepared the following tables:

PAST BIDS—COMPETITOR A'S BID VERSUS LOCKJAW COST		PAST BIDS—COMPETITOR B'S BID VERSUS LOCKJAW COST	
A's Bid (Above Lockjaw Cost)	Relative Frequency	B's Bid (Above Lockjaw Cost)	Relative Frequency
\$2,400	1/3	\$2,400	1/4
1,200	1/3	1,200	1/2
600	1/3	600	1/4

Furthermore, there was no consistent pattern between the bids of *A* and *B* (i.e., they were statistically independent). Assume that Lockjaw has only three possible bids: (1) cost + \$2400; (2) cost + \$1200; (3) cost + \$600. Which bid should be chosen? What is the expected profit?

Hint: Calculate the probability, for each alternative, of (1) winning outright, (2) tying with one competitor, and (3) tying with both competitors. Then set up profit (payoff) tables and calculate the expected profit for each strategy.

16. The Lark Company is considering replacing its No. 1 deplaning machine which is in need of considerable repair. There are two machines with which to replace it. Machine *A* is a completely automatic machine and could save Lark a considerable amount by eliminating the work that is now done manually. Machine *A* costs \$75,000.

Machine *B*, on the other hand, costs only \$20,000 and can turn out a

product of equal quality. It is only slightly more mechanized than the current machine and hence would have considerably higher labor operating costs than Machine A.

The decision about which machine to purchase hinges to a large extent upon the projected sales. But the sales manager is very uncertain about what the future sales will be. At the moment, Lark is the dominant firm in the industry. However, the sales manager thinks it is quite possible that several large manufacturers will enter the market soon. When questioned further, the sales manager stated that he believed that there was a 30 percent chance that Lark could maintain its dominant position, a 50 percent chance that it could keep a moderate share of the market, and a 20 percent chance that it would slip to a small share of the market.

Earnings were then projected for each of these possibilities, as shown in the table:

DISCOUNTED FUTURE CONTRIBUTION OF PRODUCT
(EXCLUDING THE INITIAL COST OF MACHINE)

	Share of Market		
	Dominant	Moderate	Small
Machine A	\$225,000	\$125,000	\$55,000
Machine B	120,000	80,000	45,000

Which machine should Lark buy? Why?

17. Hony Pharmaceuticals is a manufacturer engaged in the development and marketing of new drugs. The chief research chemist at Hony, Dr. Bing, has informed the president, Mr. Hony, that recent research results have indicated a possible breakthrough to a new drug with wide medical use. Dr. Bing urged an extensive research program to develop the new drug. He estimated that with expenditures of \$100,000 the new drug could be developed at the end of a year's work. When queried by Mr. Hony, Dr. Bing stated that he thought the chances were excellent, "9 or 10 to 1 odds," that the research group could in fact develop the drug.

Mr. Hony, worried about the sales prospects of a drug so costly to develop, talked to his marketing manager Mr. Margin, who said that the market for the potential new drug depended upon the acceptance of the drug by the medical profession. Margin also stated that he had heard rumors that several other firms had been considering developing such a drug. If several firms developed competing drugs they would have to split the market among them. Hony asked Margin to make future market estimates for different situations, including estimates of future profits. Margin made the estimates shown in the table:

Market Condition	Likelihood	Present Value of Profits
Large Market Potential	0.1	\$500,000
Moderate Market Potential	0.6	250,000
Low Market Potential	0.3	80,000
	1.0	

Margin pointed out that the profit figures did not include the costs of research and development or the cost of introducing the product (\$50,000). This latter cost would be incurred only if the firm decided to enter the market *after* the drug was developed.

Mr. Hony was somewhat concerned about spending the \$100,000 for development of the drug in the face of such an uncertain market. He returned to Dr. Bing and asked if there was some way to develop the drug more cheaply or to postpone development until the market position was clearer. Dr. Bing said that he would prefer his previous suggestion—an orderly research program costing \$100,000—but that an alternate was indeed possible. The alternate plan called for a low-level research program for 8 months and then a “crash” program for 4 months. The cost of this would be \$40,000 for the low level part plus \$110,000 for the crash program. Dr. Bing did not think this program would change the chances of a successful product development. One advantage of this approach, Dr. Bing added, was that the question of whether the drug could be developed successfully would be known at the end of the 8-month period. The decision could then be made at the end of 8 months on whether to undertake the crash program. When consulted, the marketing manager, Mr. Margin, stated that at the end of 8 months he would be able to estimate the market potential accurately.

Mr. Hony inquired about the possibility of waiting until other drugs were on the market and then developing a drug on the basis of a chemical analysis of the competitive drug. Dr. Bing said that this was indeed possible and that such a drug could be developed for \$50,000. Mr. Margin was dubious of the value of such an approach. He said that the first drugs out usually got the greater share of the market. He estimated that returns would only be about 40 percent of those given in the table. In addition, he indicated that there was a good chance, say 1 out of 3, that no equivalent competitive drug would be marketed—in which case Hony would have nothing upon which to develop a drug.

- a) Draw a decision tree for this problem.
- b) Which action should Mr. Hony take in order to maximize his expected profit?

SELECTED READINGS

Selected readings for this chapter are included in the list that appears on page 247.

10. DECISION-MAKING UNDER UNCERTAINTY: THE VALUE OF ADDITIONAL INFORMATION

CHAPTER 9 INTRODUCED a logical structure for decision-making in an uncertain environment. In this chapter, we wish to elaborate upon these procedures from a different point of view. This will lead to the question of whether the decision-maker should act *now* with the information available or whether he should postpone the decision and gather additional information.

OPPORTUNITY LOSS

In order to introduce the concept of opportunity loss, let us return to the example of the previous chapter. Recall that the Zip Car Rental Company leased cars for \$7 per day and rented them in turn for \$10 per day. The payoff table for the decision, including the probabilities and expected values, is shown in Table 10-1. In constructing such a table, it was important to include only real cash or "out-of-pocket" expenses and revenues. We explicitly excluded all fixed costs, as well as profits or costs from missed opportunities.¹ But these missed opportunity costs give us important insights into the decision problem.

Consider the action "lease 12 cars." If we lease 12 cars and receive only 10 rental requests, our profit is \$16. This is not the best we could have done with 10 requests, since, had we leased 10 cars, we would have made \$30. We had an *opportunity* to make 14 additional dollars, if only we had known the true number of requests. The amount \$14, then, is the *opportunity loss* associated with the decision "lease 12 cars" and the event "10 rental requests." It is the amount we fall short of the

¹ Such concepts are implicitly included in the table, as we shall see immediately.

optimal decision, given the event (in this case, 10 requests). The opportunity loss has also been designated by the term *regret*, and this term is very descriptive. If, after the fact, we rent only 10 cars, but have 12 cars available, we "regret" having leased the two extra cars and thus having lost an extra \$14 in profit.

Table 10-1
PAYOFF TABLE FOR ZIP CAR RENTALS
(DOLLARS PROFIT)

Event: Number of Rental Re- quests	Probability	Actions: Number of Cars Leased							
		10	11	12	13	14	15	16	17
10	0.05	\$30*	23	16	9	2	-5	-12	-19
11	0.05	30	33*	26	19	12	5	-2	-9
12	0.10	30	33	36*	29	22	15	8	1
13	0.15	30	33	36	39*	32	25	18	11
14	0.20	30	33	36	39	42*	35	28	21
15	0.25	30	33	36	39	42	45*	38	31
16	0.15	30	33	36	39	42	45	48*	41
17	0.05	30	33	36	39	42	45	48	51*
	1.00								
Expected Profit		30.00	32.50	34.50	35.50†	35.00	32.50	27.50	21.00

* Figure represents maximum possible profit for each event.

† Maximum expected profit.

There is an opportunity loss for each combination of event and action. We can draw up an *opportunity loss table* by subtracting each profit figure in a row from the *maximum* profit (asterisk) shown in that row. This is done in Table 10-2. Note that, in this decision situation, there are zeros on the diagonal of the table from the upper left to the lower right. This results because one can do no better than lease the exact number of cars that are requested; in each case this is the best action for the given event. There is no opportunity loss or regret. The values above the diagonal are in multiples of \$7 (the daily lease rate) representing the opportunity losses of having leased more cars than were requested. Below the diagonal, the values are in multiples of \$3, representing the profit that is forgone when there are more requests than leased cars available (\$10 revenue less \$7 cost per car).

It is important not to confuse opportunity loss with the accounting term "loss," which means a negative profit. Opportunity loss is always positive; it is measured relative to some optimal or "best" profit.

We can compute the *expected opportunity loss* in the same way as we

computed expected profit—by multiplying each opportunity loss in a given column by its probability and adding the products. This yields a weighted average of opportunity losses for each action—the loss we might expect in the long run if we consistently chose that action. Table 10-2 shows the expected opportunity loss (EOL) for each action. Note that the alternative “lease 13 cars” has the least EOL. That is, if we put in a firm order to lease 13 cars each day we would have less regret over lost

Table 10-2
OPPORTUNITY LOSS TABLE FOR ZIP CAR RENTALS
(DOLLARS REGRET)

Event: Number of Rental Re- quests	Probability	Actions: Number of Cars Leased							
		10	11	12	13	14	15	16	17
10	0.05	\$ 0	\$ 7	\$14	\$21	\$28	\$35	\$42	\$49
11	0.05	3	0	7	14	21	28	35	42
12	0.10	6	3	0	7	14	21	28	35
13	0.15	9	6	3	0	7	14	21	28
14	0.20	12	9	6	3	0	7	14	21
15	0.25	15	12	9	6	3	0	7	14
16	0.15	18	15	12	9	6	3	0	7
17	0.05	21	18	15	12	9	6	3	0
	1.00								
Expected Op- portunity Loss		\$12.00	9.50	7.50	6.50*	7.00	9.50	14.50	21.00

* Minimum expected opportunity loss.

opportunities than if we leased any other number of cars consistently. This must necessarily be the case. The use of opportunity losses is simply another way of looking at the same problem that was illustrated in Table 10-1. And that action with the highest expected profit must also have the least expected opportunity loss. That is, we can minimize EOL as our decision criterion as an alternative to maximizing expected profit.

EXPECTED VALUE OF PERFECT INFORMATION

We now turn to the problem of whether additional information should be collected before action is taken. More specifically, we would like to know how much additional profit would result from having more information. Thus, we can compare the value of this information with the cost of obtaining it.

While it often is not possible to assess the value of any specific amount of information, in terms of added profit, it is possible to put an upper limit on the value of additional information. In particular, we can determine the value of *perfect* information—that is, the exact knowledge of what event will occur.

Let us call the *expected value of perfect information* (or EVPI) the expected savings (or additional profit) from knowing the exact event that will occur. Now, the *expected value of perfect information is precisely the expected opportunity loss of the best action*. Recall that opportunity loss is the additional profit associated with picking the best decision. With perfect information about what will happen we could always make the best decision. Perfect information will save us precisely the amount of the opportunity loss. By multiplying the opportunity losses by the probabilities that each event will occur we obtain the expected opportunity loss and simultaneously the expected value of perfect information.

In the Zip Company case, the action “lease 13 cars” is the best action in the face of uncertainty about how many rentals will be needed. The opportunity losses (from Table 10–2) for this alternative are repeated in Table 10–3.

Table 10–3

OPPORTUNITY LOSSES FOR ACTION: LEASE 13 CARS

Event: Number of Rental Requests	Probability	Opportunity Loss	Expected Value
10	0.05	\$21	1.05
11	0.05	14	0.70
12	0.10	7	0.70
13	0.15	0	0
14	0.20	3	0.60
15	0.25	6	1.50
16	0.15	9	1.35
17	0.05	12	0.60
	1.00	EOL =	\$6.50

When 10 rentals are requested, there is an opportunity loss of \$21. If this event had been predicted beforehand, as it would with perfect information, the decision-maker would have saved \$21. Hence, perfect information is worth \$21 in the event “10 rental requests” occurs. If 13 rentals are requested, perfect information is worth nothing because we would be making the best decision anyway. Perfect information is, in a

sense, like a crystal ball, predicting accurately the future event. But before we have the crystal ball (i.e., perfect information) we do not know how much it will save us. It might save us \$21 or \$14 or any of the values in Table 10-3, column 3. The expected savings with the crystal ball (i.e., EVPI) is obtained by multiplying the probabilities by the savings (the opportunity loss) for each event and adding the products.

In most decision situations, it is not possible to obtain perfect predictions; accurate crystal balls just are not available. The EVPI puts an *upper limit* on what one would pay for additional information. In our example, $EVPI = \$6.50$. A system for predicting future rental requests, no matter how accurate, would be worth no more than \$6.50 per day.

Profit under Certainty: An Alternative Method for Determining EVPI

Another method for determining EVPI is to first determine the expected profit that would result if perfect information were available. Table 10-4 shows the optimal profits for each possible event. Even if

Table 10-4
PROFIT UNDER CERTAINTY

Event: Number of Rental Requests	Probability	Best Action	Profit from Best Action	Expected Value
10	0.05	lease 10 cars	\$30	\$ 1.50
11	0.05	lease 11 cars	33	1.65
12	0.10	lease 12 cars	36	3.60
13	0.15	lease 13 cars	39	5.85
14	0.20	lease 14 cars	42	8.40
15	0.25	lease 15 cars	45	11.25
16	0.15	lease 16 cars	48	7.20
17	0.05	lease 17 cars	51	2.55
Expected Profit under Certainty				42.00

we could make the best profit for each event, we do not know which will occur. Hence, we take the expected value. This is the *expected profit under certainty*, \$42.20, and measures the profit level obtainable with a perfect predictor (i.e., knowing in advance the number of cars needed each day and leasing just that number). On the other hand, our best expected profit under *uncertainty* was \$35.50, obtained by leasing 13 cars each day throughout the period. The difference between these numbers is \$6.50; this is the expected value of the perfect information (EVPI).

An Example

A manufacturer must decide whether to build a new plant. The profitability of the plant will depend upon future economic conditions (either stability or growth). The payoffs for various actions and events and the subjective probabilities that the manufacturer assigns to stability and growth are shown in Table 10-5.

Table 10-5
PAYOFF TABLE
PROFITS FROM BUILDING NEW PLANT
(MILLIONS OF DOLLARS)

Event: Level of National Economy	Probability	Actions	
		Build	Do Not Build
Stability.....	0.2	3	5
Growth.....	0.8	16	12
	1.0		
Expected Profit		13.4	10.6

The opportunity loss table for this problem is shown as Table 10-6.

Table 10-6
OPPORTUNITY LOSS TABLE
(MILLIONS OF DOLLARS)

Event: Level of National Economy	Probability	Actions	
		Build	Do Not Build
Stability.....	0.2	2	0
Growth.....	0.8	0	4
	1.0		
Expected Opportunity Loss		0.4	3.2

If the economy is stable, "do not build" is the better action and hence has an opportunity loss of zero. If instead the plant *were* to be built, it would reduce profit by \$2 million, relative to the best alternative. Hence, the opportunity loss of "build," if stability occurs, is \$2 million.

Similarly, if there is to be economic growth, "build" is the best alternative and has zero regret (opportunity loss). If the decision-maker failed to build and there is growth, his opportunity loss would be \$4

million since his profit would be reduced by this much relative to the optimal decision.

The expected value of perfect information is equal to the EOL of the best decision. In this case, the best decision is "build" and $EVPI = 0.4$ million or \$400,000.

Alternatively, we can calculate the profit under certainty as shown in Table 10-7. EVPI is then determined as the expected profit under certainty less the profit under uncertainty ($13.8 - 13.4 = 0.4$), yielding 0.4 million, as above.

Table 10-7
CALCULATION OF EXPECTED PROFIT UNDER CERTAINTY
(MILLIONS OF DOLLARS)

Event: Level of National Economy	Prob- ability	Best Action	Profit from Best Action	Expected Value
Stability	0.2	Build	5	1.0
Growth	0.8	Do not build	16	12.8
Expected Profit under Certainty				13.8

Since this is a sizable amount, the decision-maker might profitably seek more information about future economic trends before making his decision. This is not to say that one could ever get perfect information on future events. Perhaps the decision-maker could hedge somewhat in this case by proceeding with the plans but still keeping alive the possibility that the project might be canceled if economic growth did not justify it.

LINEAR PROFIT FUNCTIONS

In the previous chapter and in the earlier sections of this chapter we presented a general framework for decision-making under uncertainty. In the remainder of this chapter we shall present some special cases in which the analysis is considerably simpler than heretofore. Fortunately, these cases encompass many decision problems and have rather broad usefulness.

The first such instance occurs when the profit for a given action can be represented as a linear function of the unknown variable. Let us illustrate this.

A manufacturer of children's toys has a new toy which he is considering marketing nationwide. The toy is a novelty item which would be discontinued after a single national selling campaign. The variable cost

to manufacture the toy is 12 cents. The selling price to retail outlets is 57 cents, so the unit profit is $\$.57 - \$.12 = \$.45$. A national advertising campaign to sell the product would cost \$2.7 million. Management is uncertain about how many of the toys will be sold. The probability distribution assigned to the unknown variable—number of units sold—appears in Table 10-8. The possible actions are (1) market the new product or (2) abandon the product.

Table 10-8
PROBABILITIES AND EXPECTED VALUES OF
TOY SALES

Event: No. Sold (Million), X	Probability, $P(X)$	Expected Value (Millions of Units), $X \cdot P(X)$
4	0.2	0.8
6	0.3	1.8
8	0.4	3.2
10	0.1	1.0
	1.0	$E(X) = 6.8$

We could, of course, analyze this problem by drawing up a payoff table and proceeding as outlined in Chapter 9 and the first part of this chapter. Instead, let us find an equation that will relate profit to the unknown number of items sold (X). There is one equation for each action:

Market the product: Profit (dollars) $\pi = -2,700,000 + 0.45X$

Abandon the product: Profit = 0

These equations are graphed in Chart 10-1.

The first equation contains a negative \$2.7 million (the cost of promotion campaign) and a variable contribution per unit of 45 cents times the number of units sold. Thus, if 8 million were sold, profit would be:

$$\pi = -2,700,000 + (0.45)(8,000,000) = +\$900,000$$

Note that the profit equations are linear. That is, they are of the form

$$\pi = a + bX \quad (1)$$

where π = profit; a and b are constants; and X is the unknown variable. When this is the case, the expected profit, $E(\pi)$, can be found by the following equation:²

² This can be shown as follows: $E(\pi) = \sum P(X)\pi = \sum P(X)[a + bX] = \sum aP(X) + \sum bXP(X) = a\sum P(X) + b\sum XP(X)$. But $\sum P(X) = 1$ because $P(X)$ is a probability function and $\sum XP(X)$ is defined to be $E(X)$. Hence, $E(\pi) = a + bE(X)$, as shown on the following page.

$$E(\pi) = a + bE(X) \quad (2)$$

where $E(X)$ is the expected value of the unknown variable X .

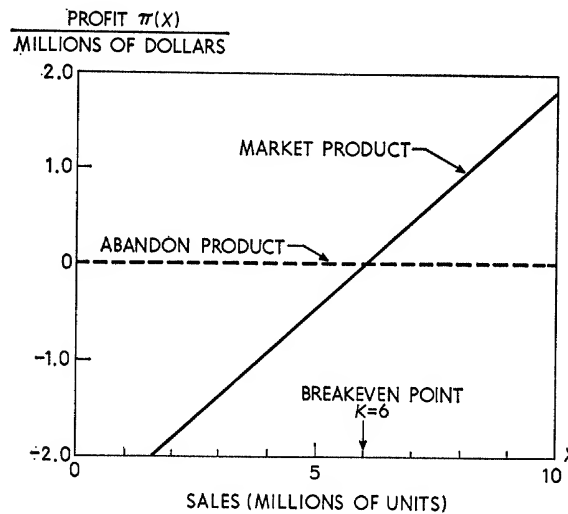
For the decision "market the product," $a = -\$2,700,000$ and $b = \$0.45$. $E(X) = 6.8$ million unit sales, as in Table 10-8. Hence, the expected profit (using Equation 2) is

$$E(\pi) = -2,700,000 + (.45)(6,800,000) = \$306,000$$

For the decision "abandon the product," both a and b are 0 and $E(\pi) = 0$. If the toy manufacturer were to act now, therefore, he would market the product, since this action has a higher expected profit than the alternative (which has zero profit).

Chart 10-1

PROFIT FUNCTIONS OF TWO ACTIONS
IN MARKETING NEW TOY



It is also instructive to calculate the "break-even" level of sales; that is, the volume of sales at which the decision-maker is indifferent between the two alternatives. In this case it is the sales necessary to cover the advertising expenses. Let us denote this break-even value by K . Then

$$\begin{aligned} \$0.45 K &= \$2,700,000 \\ K &= 6,000,000 \text{ units} \end{aligned}$$

Once this value is known, the decision-maker can simply compare the expected sales $E(X)$ with the break-even point K . If $E(X)$ is greater than K , then marketing the product will be more profitable. If $E(X)$ is

less than K , marketing the product would lead to negative profits, and it would be better to abandon the project.

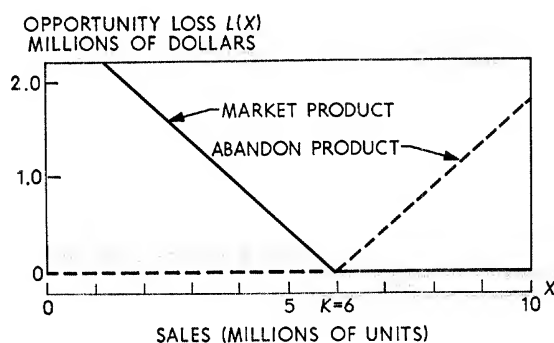
Opportunity Loss Functions

When the profit function is linear, each function describing the possible opportunity losses from a given action can be described by two connected straight lines.³ The loss functions for our illustration are shown in Chart 10-2. These functions are

$$\begin{aligned} \text{Action: Market the product} \\ \text{Opportunity loss} = L(X) &= 0 && \text{if } X \geq 6 \text{ million} \\ &\text{or } L(X) = \$0.45(6,000,000 - X) && \text{if } X < 6 \text{ million} \end{aligned}$$

$$\begin{aligned} \text{Action: Abandon the product} \\ \text{Opportunity loss} = L(X) &= \$0.45(X - 6,000,000) && \text{if } X > 6 \text{ million} \\ &\text{or } L(X) = 0 && \text{if } X \leq 6 \text{ million} \end{aligned}$$

Chart 10-2
OPPORTUNITY LOSS FUNCTIONS FOR TWO ACTIONS
IN MARKETING NEW TOY



Note that the break-even point, $K = 6$ million units, plays a key part in determining the loss functions. Their meaning is as follows: If we market the product and sales exceed the break-even value (6 million), then there is no opportunity loss, since we have made the correct decision. If, on the other hand, sales are below 6 million, our regret (loss) is 45 cents for every unit that sales fall below 6 million, since,

³ We are describing here the loss functions for two-action problems (i.e., only two actions are considered). For multi-action problems, each loss function still consists of connected straight lines, but the subsequent analysis is more complicated.

had we abandoned the project, we could have avoided this loss. Similarly, if we abandon the project and sales are at or below the break-even value, then our loss is zero, since we acted optimally. However, if sales are above 6 million, we suffer an opportunity loss of 45 cents for every unit above 6 million, since this is profit we could have obtained, had we acted optimally.

Because these loss functions are broken rather than continuous straight lines, it is not generally possible to obtain a simple expression for the expected opportunity loss (EOL) and EVPI, except in the special case of the normal distribution considered below.

However, we can compute the expected value of perfect information in our usual fashion. This is done in Table 10-9. The expected opportunity loss for the best decision is \$180,000. This is the expected value of perfect information.

Table 10-9

OPPORTUNITY LOSSES AND EXPECTED VALUE OF PERFECT INFORMATION

Event: Sales, Millions of Units, X	Probability $P(X)$	Opportunity Losses (Millions of Dollars)		Expected Value (Millions of Dollars)	
		Market Product	Abandon Product	Market Product	Abandon Product
4	0.2	\$0.9	\$0	\$0.18	\$0
6	0.3	0	0	0	0
8	0.4	0	0.9	0	0.36
10	0.1	0	1.8	0	0.18
	1.0			EOL = \$0.18	\$0.54

THE NORMAL DISTRIBUTION IN DECISION-MAKING

In making decisions under uncertainty, the decision-maker can express his subjective feelings about the unknown variable as a probability distribution. In many situations it is reasonable to use the normal distribution for this purpose. The choice of the normal distribution as a decision-making or betting distribution implies that the decision-maker feels that some value of the unknown variable is the most likely (the mean μ of the distribution); that the variable is more likely to be close to this guess than far away (the area of the normal distribution is clustered around μ); and that the unknown variable could as likely be on either side (high or low) of this guess (the normal distribution is symmetrical about μ).

The normal distribution has two parameters, μ the mean and σ the standard deviation. In finding appropriate values for these parameters to

use in his particular situation, the decision-maker must phrase some questions for himself. In order to estimate the mean μ , he must find the middle point of his betting distribution. He should be willing to bet that the unknown variable X is as likely to fall above as below μ , as in the normal distribution. Also, since two thirds of the area of the normal curve lies within one standard deviation of the mean, the decision-maker should specify a range about μ such that there is a two-thirds chance that X would be in this interval.⁴ That is, the decision-maker should estimate the value of σ such that he would be willing to bet that X will fall in the interval $\mu \pm \sigma$ with odds of 2 out of 3.

Before using this normal distribution, the decision-maker should graph it and check the probabilities implied by the distribution against his judgment. For example, he should judge the odds to be about 95 out of 100 that X will fall in the interval $\mu \pm 2\sigma$.⁵

Opportunity Loss and the Normal Distribution

When the profit for a given action can be expressed as a linear function ($\pi = a + bX$), the expected profit is also a linear function of $E(X)$; that is, $E(\pi) = a + bE(X)$ regardless of whether the decision distribution is normal or any other shape. However, when the decision distribution is normal, the expected opportunity loss can be expressed in a simplified form. Consider Chart 10-3. Here a normal distribution is superimposed upon a loss function for a given action. The expected loss is simply the probability function times the loss function summed over the whole area. The simplified formula for expected value of perfect information (the EOL of the best action) in this case is

$$EVPI = t\sigma L_N(D) \quad (3)$$

where

$$D = \left| \frac{K - \mu}{\sigma} \right| \quad (4)$$

In the above formulas, t is the slope of the opportunity loss function; μ and σ are the parameters of the normal decision distribution; K is the

⁴ An alternative procedure is to specify a symmetric interval about μ (e.g., $\mu \pm Q$, the quartile deviation) such that there is an even chance for X to be in this interval. Then $Q = 2/3\sigma$ or $\sigma = 3/2Q$. This follows from the fact that the normal distribution has about half its area in the interval $\mu \pm 2/3\sigma$.

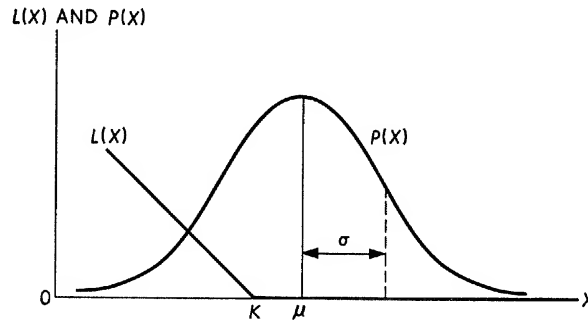
⁵ The normal distribution is at best an approximation to one's betting distribution. This distribution is continuous, whereas most decision-making distributions are discrete (e.g., sales are in integer units). Also, the normal distribution has tails that go out in both directions indefinitely, though the probabilities in these tails are quite small. Generally, we would like to truncate our decision distribution at certain points (e.g., sales cannot be negative so the probabilities of negative sales should be zero). Despite these minor inconsistencies, the normal distribution is quite adequate for many situations.

break-even point; and $L_N(D)$ is the unit normal loss function which is found by looking up D in Appendix E. The symbol $||$ means absolute value (ignoring a negative sign).

An Example. A distributor has an opportunity to market his product in a new territory. The fixed cost of this action is \$4,000 for advertising, facilities, etc. For each unit sold the distributor will make a profit of \$.10. It will thus take sales of 40,000 units to break even ($K = 40,000$).

Chart 10-3

OPPORTUNITY LOSS FUNCTION $L(X)$
AND NORMAL DISTRIBUTION $P(X)$



The distributor is something of a mathematician, but he is quite uncertain about how many units he will actually sell. He is willing to represent his uncertainty about sales by a normal distribution. Suppose that he feels that sales are as likely to be above 50,000 as below 50,000 (that is, $\mu = 50,000$). Further, suppose he assigns a probability of two thirds to the possibility that actual sales will be in the range of 25,000 to 75,000. Since this range is $50,000$ (or μ) $\pm 25,000$, the standard deviation $\sigma = 25,000$, and when presented with Chart 10-4, the decision-maker agrees that this adequately represents his betting distribution.

The profit functions are

$$\text{Open new territory: } \pi = -\$4,000 + (0.10)X$$

$$\text{Do not open new territory: } \pi = 0$$

where X is the number of units sold.

The expected profits are

$$\begin{aligned} \text{Open new territory: } E(\pi) &= -\$4,000 + \$(0.10)(50,000) \\ &= \$10,000 \end{aligned}$$

$$\text{Do not open new territory: } E(\pi) = 0$$

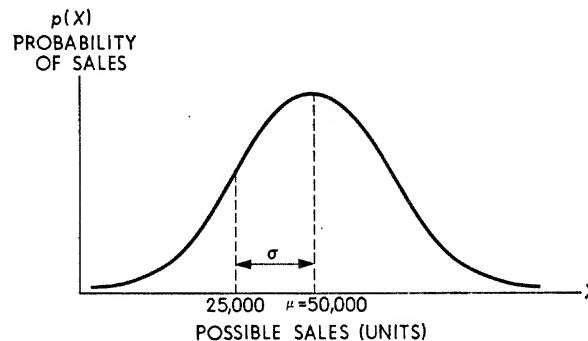
And so, with this information the decision-maker should market in the new territory.

The opportunity loss function for this optimal decision is

$$\begin{aligned} \text{Opportunity loss} = L(X) &= 0 && \text{if } X \geq 40,000 \\ \text{or } L(X) &= \$ (0.10)(40,000 - X) && \text{if } X < 40,000 \\ &= \$4,000 - (0.10)X \end{aligned}$$

Chart 10-4

NORMAL DECISION DISTRIBUTION FOR
POSSIBLE SALES IN NEW TERRITORY



Using Equations 3 and 4, we can determine the expected opportunity loss for this decision (which is EVPI, since it is the optimal decision):

$$\begin{aligned} D &= \left| \frac{K - \mu}{\sigma} \right| = \left| \frac{40,000 - 50,000}{25,000} \right| = 0.40 \\ \text{EVPI} &= t\sigma L_N(D) = (0.10)(25,000)L_N(0.40) \\ &= (0.10)(25,000)(0.2304) = \$576 \end{aligned}$$

In the above equations, the values of $\mu = 50,000$ and $\sigma = 25,000$ represent the decision-maker's normal betting distribution. The break-even sales value is $K = 40,000$ units. The slope of the loss function is $t = 0.10$; this is the loss for each unit below the 40,000 break-even level. And, finally, the value of $L_N(D) = L_N(0.40)$ is obtained from Appendix E.

Interpretation of EVPI. In the above example the expected value of perfect information is \$576. This means that the distributor would pay no more than this amount for information about his exact sales. The information he can get (studies of income, market potential, etc.) is worth a good deal less than \$576, since such information cannot give an exact prediction.

If we reexamine formulas 3 and 4, we can see what factors influence the value of EVPI.

$$EVPI = t\sigma L_N(D) \quad (3)$$

$$D = \left| \frac{K - \mu}{\sigma} \right| \quad (4)$$

Note the following: (a) The farther the break-even point (K) is from the expected sales (μ), the larger is D and the smaller are $L_N(D)$ (see Appendix E) and EVPI. Clearly, if the break-even point is well above or below the expected sales, the decision is relatively certain and additional information is of little value. On the other hand, if $(K - \mu)$ is small, even a little additional information may change the decision, and hence may be valuable. (b) The larger σ , the larger is EVPI. The standard deviation σ is a measure of the degree of uncertainty in the decision situation. The more the uncertainty, the more valuable the perfect information. (c) The symbol t represents the per unit opportunity loss. Hence, the larger t , the larger is EVPI. If t is small, the economic consequences of making the wrong decision are not serious. If t is large, they may be.

Another way of looking at EVPI is as the maximum price the decision-maker might pay for insurance to guarantee him against a loss.⁶ In the distributor example, the decision-maker should be willing to pay an insurance premium up to \$576. The insurance policy would pay the difference between the revenue from the new territory (\$.10 times the number of units sold) and the fee of \$4,000 if revenue were less than \$4,000.

Another Example. One final example will serve to review the concepts of decision-making under uncertainty presented in this chapter and the last. A manufacturer must replace some machinery that has worn out. There are two alternative types of machinery that can be

⁶ Or to guarantee him a profit if he decides not to act, when in fact a profit could have been made. In other words, the insurance would pay the opportunity loss. As a practical example of such a situation, consider the following from a front-page article in *The Wall Street Journal* of December 6, 1966:

"Good Weather, Inc., a Long Island insurance agency that specializes in unusual risks, says that for the past six years a major maker of candy has bought a policy insuring against rain or snow on Valentine's Day. Henry Fox, president of the agency, says, 'Since candy is an impulse-type purchase, the company's retail stores would be left with a large stock if the weather was bad. But people after Valentine's Day won't buy candy in heart-shaped boxes because they're afraid it might be stale. So we insure the manufacturer against the expense of transferring the candy to regular boxes.'

"The policy is for almost \$250,000, and the premium is \$10,000. It covers various cities in the Northeast and the complex payout formula is based upon the amount of snow or rain and the number of hours that it snows or rains. But so far, says Mr. Fox, he hasn't had to pay out a cent on the policy."

picked to replace the worn-out equipment. Machinery type *A* is conventional; it costs \$200,000, and has a variable operating cost of \$12 per hour (direct labor, maintenance, etc.). Machinery type *B* is largely automated; it costs \$400,000, but has a variable operating cost of only \$7 per hour. Both machines produce the same output per hour in quantity and quality.

Because of economic factors, the market for the product is in a state of flux. Hence, the required number of hours of operating time on the machinery is uncertain. Management expressed this uncertainty in terms of a normal distribution with mean $\mu = 50,000$ hours.⁷

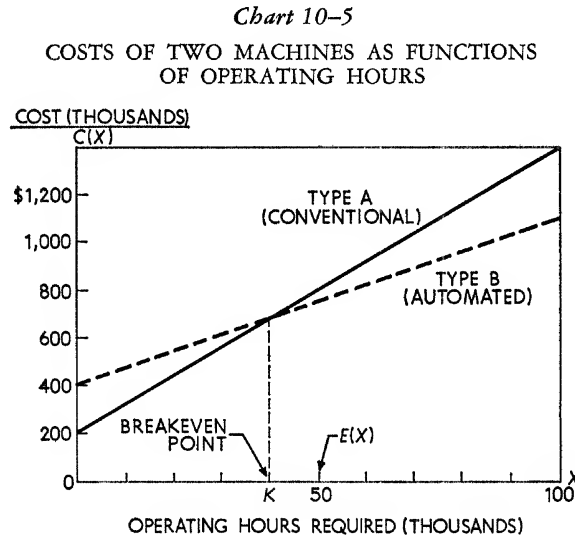
The cost functions for the two alternatives are

$$\text{Machinery Type A: Cost} = C(X) = \$200,000 + \$12X$$

$$\text{Machinery Type B: Cost} = C(X) = \$400,000 + \$7X$$

where X is the actual number of machine hours used.

The cost functions are graphed in Chart 10-5. Note that by setting the equations equal to each other, and solving for X , the break-even



point (when the two machinery types have the same cost) occurs at 40,000 hours. If less than 40,000 operating hours are required, the conventional machinery (Type *A*) has less cost. For more than 40,000

⁷ Since these hours probably would be spread over several years, discounting procedures are appropriate. Further, there are tax factors associated with depreciation that are relevant to the decision. We have omitted these factors in order to concentrate on the decision analysis. See the reference to Harlan, Christenson, and Vancil (pp. 239-65) at end of this chapter for a discussion of these topics.

hours, the automated machinery (Type B) has the cost advantage. And since the expected number of hours $E(X) = 50,000$, the purchase of the Type B machinery is the optimal decision.

The same conclusion can be reached by determining the *expected cost* for the choice of each machine:

$$\text{Type A: } E(C) = \$200,000 + \$12(50,000) = \$800,000$$

$$\text{Type B: } E(C) = \$400,000 + \$7(50,000) = \$750,000$$

Machinery Type B has \$50,000 expected cost less than its alternative.

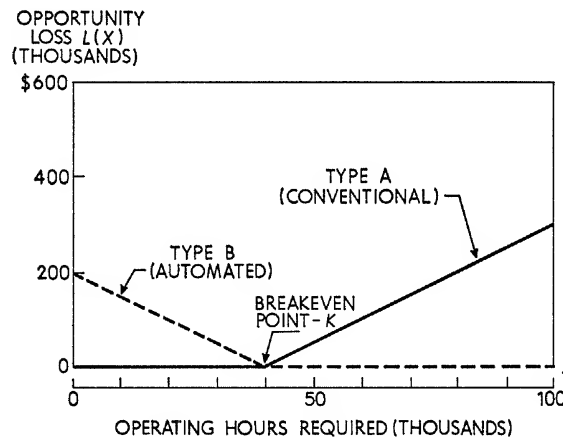
The *opportunity loss* functions are

$$\begin{aligned} \text{Type A: } L(X) &= \$5(X - 40,000) = \$5X - \$200,000 && \text{if } X > 40,000 \\ \text{or } L(X) &= 0 && \text{if } X \leq 40,000 \end{aligned}$$

$$\begin{aligned} \text{Type B: } L(X) &= 0 && \text{if } X \geq 40,000 \\ \text{or } L(X) &= \$5(40,000 - X) = \$200,000 - 5X && \text{if } X < 40,000 \end{aligned}$$

They are graphed as Chart 10-6.

Chart 10-6
OPPORTUNITY LOSS FUNCTIONS
FOR TWO MACHINES



In the above functions, the break-even point K is 40,000 hours. The slope t of the nonzero opportunity loss functions is \$5 (or $-\$5$ for Machinery Type B). This needs some explanation. The \$5 is the *difference* between the variable operating costs of the two types of machinery ($\$12 - \$7 = \$5$).⁸ If Machinery Type B is purchased, and hours

⁸ In two-action problems the slope of the nonzero parts of the loss function is always the difference between the slopes of the profit or cost functions. In the previous examples the slope of one of the profit functions was zero, so that we did not have to make this point.

required fall below 40,000, the manufacturer incurs costs of \$5 per hour for each hour under 40,000 in excess of what he would have incurred if he had acted optimally.

The expected value of perfect information is

$$\begin{aligned} \text{EVPI} &= t\sigma L_N(D) \quad \text{where } D = \left| \frac{K - \mu}{\sigma} \right| \\ D &= \left| \frac{40,000 - 50,000}{20,000} \right| = 0.50 \\ \text{EVPI} &= (\$5)(20,000)L_N(0.50) = (\$100,000)(0.1978) \\ &= \$19,780 \end{aligned}$$

SUMMARY

The previous chapter introduced methods for decision-making under uncertainty by which we could answer the following question: "If we must act now with the information available, what is the optimal act?" The first part of this chapter was directed at the question: "Should we act now or postpone the decision and collect additional information before acting?"

We first consider *opportunity loss*, which is part of the world of "might have been." It is the difference between the profit actually achieved and the profit that could have been obtained had the optimal action for a given event been selected. An *opportunity loss table* shows the opportunity loss for each combination of action and event. The *expected opportunity loss* (EOL) of any action is then the weighted average of the opportunity losses associated with that action, the weights being the probabilities of the various events.

The *expected value of perfect information* (EVPI) is the additional profit that could be made if the decision-maker knew beforehand and could pick the optimal action for every possible event. The expected opportunity loss (EOL) of the best action is uniquely the expected value of perfect information (EVPI). The expected value of perfect information can also be obtained by calculating the *expected profit under certainty* and subtracting the highest expected profit under uncertainty. The expected value of perfect information is an important concept in the decision whether to act now or later. If EVPI is small, it means that our uncertainty is small when measured in economic terms; hence, there is little to be gained from additional information. On the other hand, if EVPI is large, then there is room for considerable improvement in the available information; possibly we should seek more information before acting. However, since most obtainable information

is not a perfect predictor, we generally cannot place a specific value on the information; we can only place an upper limit on its value.

When the profit for a given action can be expressed as a *linear* function of the unknown variable, then the expected profit for that action can be determined simply from the expected value of the unknown variable. The opportunity loss function is composed of two linear pieces.

The use of the normal distribution as a decision-making or betting distribution implies symmetrical, unimodal shaped distribution, with the probability clustered near the center.

Under certain conditions—a two-action problem, linear profit functions, and a normal betting distribution—EVPI can be expressed as a simple formula. In this instance, EVPI depends *directly* upon the standard deviation of the betting distribution and on the per unit opportunity loss; EVPI depends *inversely* upon the distance of the break-even point from the mean of the betting distribution.

One way to obtain information in decision situations is to take a sample. In Chapters 15 and 16 we consider the value of information obtained from a sample.

FORMULAS

Linear profit function: $\pi = a + bX$

Expected profit for linear profit function: $E(\pi) = a + bE(X)$

Expected value of perfect information for two-action problems with normal betting distribution and linear profit functions:

$$EVPI = t\sigma L_N(D) \quad \text{where } D = \left| \frac{K - \mu}{\sigma} \right|$$

PROBLEMS

1. Refer to Problem 3 of Chapter 9.
 - a) Prepare an opportunity loss table for this decision situation.
 - b) Calculate the expected opportunity loss for each action.
 - c) What is EVPI?
 - d) What is the expected profit under certainty?
2. Refer to Problem 6 of Chapter 9.
 - a) Prepare an opportunity loss table.
 - b) What is EVPI? Explain its meaning in this decision situation.
3. Refer to Problem 7 of Chapter 9.
 - a) Prepare an opportunity loss table.
 - b) What is the expected profit under certainty?
 - c) What is EVPI?

4. Refer to Problem 9 of Chapter 9.
 - a) Draw up an opportunity loss table for the two actions: (1) do not take the lease; (2) drill without test.
 - b) What is EVPI, assuming that Gusher feels that there are two chances out of five that oil is present?
5. Refer to Problem 4 above. Suppose, after leasing the land, Gusher can conduct certain geological tests to ascertain the presence of oil. The geological tests are not error-free, however. If oil is present, the tests have an 80 percent chance of indicating so. When oil is not present, the tests will still indicate oil 30 percent of the time. The tests cost \$20,000. Suppose that Gusher, independent of the tests, feels that there are 2 chances out of 5 that oil is present.

Draw up a payoff table. Which alternative should Gusher take: turn down the lease, test and decide to drill on the basis of the test, or drill without the test?
6. Refer to Problem 10 of Chapter 9.
 - a) What is the expected value of perfect information in this decision situation?
 - b) How might the decision-maker obtain additional information?
7. Refer to Problem 11 of Chapter 9.
 - a) Determine the EOL of each action.
 - b) Do you think IJK should obtain additional information about the financial position of new customers such as Lastco? Suppose, for example, a credit rating company could give an opinion of a potential customer for a \$200 fee.
 - c) Suppose the fee of the credit rating company was only \$50. And, based upon past experience, the ratings (good, medium, poor) related to IJK experience as follows:

CREDIT RATINGS BY CUSTOMER CLASSIFICATION
(PERCENT OF TOTAL)

Credit Evaluation Rating	Event			
	Failed	Financial Troubles	Sporadic Customer	Good Customer
Good.....	0%	10%	40%	40%
Medium.....	40	50	50	50
Poor.....	60	40	10	10
Total	100	100	100	100

Would it be worthwhile to use the credit rating company to help screen customers?

8. Refer to Problem 16 of Chapter 9. The president of Lark suggests that the decision be postponed a year. He notes the great degree of uncertainty about the future market position of the Lark Company, and he feels that the situation will be somewhat clearer in a year. It is determined that the pres-

- ent No. 1 deplaning machine can be repaired to last a year for \$3,000. Should the decision be postponed? Why or why not?
9. Refer to Problem 14 of Chapter 9. Express the profits for each action as a linear function. Calculate expected sales. Use this value to determine the expected profit for each action.
10. The quality of a manufactured product varies from day to day due to weather conditions, machine settings, worker productivity, and other factors. Over the past 100 days the quality (fraction of the items which were defective) had the following frequency distribution:

<i>Quality (Fraction of Items Defective)</i>	<i>Relative Frequency</i>
0.01	0.20
0.02	0.40
0.04	0.40
0.06	0.10
0.08	0.05
0.10	0.03
0.15	<u>0.02</u>
	1.00

- a) Using past relative frequencies as probabilities, what is the expected fraction defective?
- b) Suppose that each defective item causes rework costs of \$1.50 when the item is included in a final assembly operation. Express the rework costs for a lot of 1,000 items as a function of the fraction defective.
- c) Use the answers to *a* and *b* to determine the expected rework cost.
11. The Zippy Razor Company makes a contribution (price minus variable cost) of 8 cents on each package of razor blades sold. Fixed costs of operating (costs independent of the sales level) are \$180,000. The following probabilities are assigned to various sales levels for next year.

<i>Sales (Thousands of Packages)</i>	<i>Probability</i>
100	0.05
150	0.05
200	0.10
250	0.40
300	0.30
350	<u>0.10</u>
	1.00

- a) Express Zippy profits as a function of sales.
- b) What is the break-even point?
- c) Calculate expected sales and use this value to determine expected profit.
12. In *a* through *d* below, calculate EVPI, using the indicated values of the mean μ and standard deviation σ of the normal betting distribution; the break-even value K ; and the slope of the loss function t .

- a) $\mu = 50, \sigma = 20, K = 60, t = 100.$
- b) $\mu = 100, \sigma = 25, K = 80, t = 15.$
- c) $\mu = 15, \sigma = 5, K = 25, t = 1.0.$
- d) $\mu = 85, \sigma = 15, K = 85, t = 40.$

13. In *a* through *d* below, calculate EVPI, using the indicated values of the mean μ and standard deviation σ of the normal betting distribution; the break-even value K ; and the slope of the loss function t .

- a) $\mu = 100, \sigma = 40, K = 160, t = 0.5.$
- b) $\mu = 65, \sigma = 15, K = 50, t = 60.$
- c) $\mu = 45, \sigma = 20, K = 50, t = 0.005.$
- d) $\mu = 120, \sigma = 30, K = 110, t = 1.0.$

14. The GHK Company was considering a new advertising campaign to increase sales. The advertising would cost \$100,000. Management felt that the campaign would "most probably" increase sales by \$1 million, but there was considerable uncertainty about this figure. When pressed for more details, the GHK management said that there was about 1 chance in 3 that the sales increase would be outside the range 0.8 to 1.2 million dollars. GHK was then making a net profit of 12 percent on total sales.
- a) State the profit functions for the two actions: (1) advertise; (2) do not advertise.
 - b) State the opportunity loss functions for the two actions. What is the break-even point?
 - c) Draw a normal distribution describing GHK sales estimates.
 - d) Which action should GHK choose? What is EVPI?
15. The Flavor Coffee Company was considering the use of a new type of can to package its coffee. The president felt that the new can would have appeal to consumers and would increase sales. In fact, he was willing to bet, with even odds, that sales over the next three years would be increased at least 2 million pounds. Furthermore, he was willing to bet, again with even odds, that the sales increase would be in the range 1.5 to 2.5 million pounds. Flavor currently makes a net profit (price minus variable cost) of 12 cents on a pound of coffee. The cost of change over to the new can, however, is large—about \$200,000 in costs of new machinery, etc.
- a) Express profits from the new can as a function of sales. What is the break-even point?
 - b) State the opportunity loss functions for the actions: (1) use the new can; (2) do not use the new can.
 - c) Draw a normal distribution describing Flavor sales estimates.
 - d) If Flavor must act now, what action should be taken? What is EVPI?
16. The Central Cities Electric Company had a problem with pole fires and service failures due to dust collecting on high voltage insulators. When the first heavy fog or light rain of the season came, the dust became a conductor of electricity, with resulting pole fires and short circuits. To combat this problem, Central Cities periodically washed the insulators with pure water from a high-pressure hose.

One problem centered upon when to wash the insulators. During the summer and early fall, it did not rain in the Central Cities area. The first fog or light rain of the season was an uncertain event. From the weather records for the fifty years 1917–1966, the economic analysis section of Central Cities was able to obtain the following data:

<i>First Light Rain or Fog Occurred During Week of</i>	<i>Number of Times (i.e., Frequency) in Period 1917–66</i>
September—1st week.....	1
2nd week.....	1
3rd week.....	2
4th week.....	1
October 1st week.....	4
2nd week.....	2
3rd week.....	12
4th week.....	5
5th week.....	10
November 1st week.....	6
2nd week.....	3
3rd week.....	2
4th week.....	1
	<hr/> 50

As a matter of policy, all insulators on high-voltage power lines were washed in the last week in August. The question arose about whether the insulators should be washed again, and if so, when.

The economic analysis section performed a study relating the amount of accumulated dust (measured in number of weeks' worth) to costs of pole fires and disrupted service. This study showed:

<i>Number of Weeks Accumulated Dust Before Fog or Rain</i>	<i>Cost of Pole Fires and Disrupted Service, Thousands of Dollars</i>
1	2
2	4
3	10
4	18
5	24
6	30
7	35
8	39
9	42
10	44
11	45
12	45
13	45

Assume that it costs \$12,000 to wash the insulators on all high-voltage power lines; assume also that once fog or rain occurs, it rains sufficiently often thereafter that there are no losses from fires and disrupted services due to dust on the insulators.

- a) Should Central Cities wash the insulators again? If so, when should it be done? Assume that the washing must be scheduled by September 1 (i.e., it is not possible to decide on a week by week basis whether or not to wash).

- b) What is the expected value of perfect information in this decision situation?
 - c) Should the insulators be washed twice (in addition to the August washing) during a season?
 - d) Suppose that the decision about washing the insulators could be made on a week by week basis. How would your analysis change? Discuss briefly.
17. The ABC Company is trying to decide which machine to purchase to produce a new product. There are two alternatives:

Machine 1: Purchase Cost = \$120,000
Direct variable cost = \$2.00 per unit
Machine 2: Purchase Cost = \$350,000
Direct variable cost = \$1.00 per unit

Management was undecided about which machine to purchase because of considerable uncertainty about the sales level that would be attained by the new product. Most pessimistic estimates ranged as low as 40,000 units per year while most optimistic estimates were as high as 120,000 units per year. Management felt that 80,000 units per year was perhaps the most probable forecast. Furthermore, they felt that the odds were 2 out of 3 that sales would be somewhere between 60,000 and 100,000. The investment was to be considered over a period of 3 years.

- a) Determine the sales level at which management would be indifferent as to which machine to purchase.
 - b) Based only on the above information, which machine should be purchased?
 - c) What is the expected value of perfect information?
18. Refer to the quotation from *The Wall Street Journal* contained in footnote 6 on page 238. Comment on the decision by the candy manufacturer to buy the insurance and pay the \$10,000 premium from the point of view of:
- a) the expected value of perfect information
 - b) the decision-maker's utility curve for money.

SELECTED READINGS

BIERMAN, H.; BONINI, C.; FOURAKER, L.; and JAEDICKE, R. *Quantitative Analysis for Business Decisions*, Homewood, Illinois: Richard D. Irwin, 1965.

Chapters 3, 4, and 7 treat payoff tables, opportunity loss tables, expected value of perfect information, linear profit functions, normal probability distributions, and expected value of perfect information for the normal distribution case.

CHERNOFF, HERMAN, and MOSES, LINCOLN. *Elementary Decision Theory*. New York: John Wiley, 1959.

A detailed treatment of decision-making in the face of uncertainty.

CLELLAND, RICHARD C.; DE CANI, JOHN S.; BROWN, FRANCIS E.; BURSK, J. PARKER; and MURRAY, DONALD S. *Basic Statistics with Business Applications*. New York: John Wiley, 1966.

Chapter 15 is a general treatment of decision-making under uncertainty, including a discussion of minimax and Bayes strategies.

HADLEY, G. *Introduction to Probability and Statistical Decision Theory*. San Francisco: Holden-Day, 1967.

This book presents detailed treatment of the subjects in Chapters 7 through 10 (and 15–16) of this text. It includes material at both introductory and advanced levels.

HARLAN, N.; CHRISTENSON, C.; and VANCIL, R. *Managerial Economics: Text and Cases*. Homewood, Illinois: Richard D. Irwin, 1962.

Section III presents text and several short cases relevant to decision-making under uncertainty.

LUCE, R. DUNCAN, and RAIFFA, HOWARD. *Games and Decisions*. New York: John Wiley, 1957.

Chapter 2 is a good presentation of the role of utility in decision-making.

Chapter 13 compares different decision criteria in the face of uncertainty.

MAGEE, JOHN F. "Decision Trees for Decision-Making," in *Harvard Business Review* (July–August 1964) and "How to Use Decision Trees in Capital Investment," *Harvard Business Review* (September–October 1964).

These two articles describe the basic ideas about decision trees and show their application to several types of management decision problems.

SCHLAIFER, ROBERT. *Introduction to Statistics for Business Decisions*. New York: McGraw-Hill, 1961.

Chapters 1, 2, and 4 present the basic elements for decision-making under uncertainty. Chapter 20 treats decision-making with the normal distribution.

TEICHROEW, DANIEL. *Introduction to Management Science: Deterministic Models*. New York: John Wiley, 1964.

This book is concerned with models for decision-making under certainty.

II. INTRODUCTION TO STATISTICAL INFERENCE

THE ABILITY TO MAKE valid generalizations and predictions from sample data is an important step forward in scientific knowledge. The methods of collecting data from samples were described in Chapter 2. Chapters 4 to 6 presented the necessary tools of analysis—frequency distributions, averages, and measures of dispersion. Chapters 7 and 8 developed the fundamentals of probability theory. These basic concepts can now be brought together in the study of statistical inference.

Statistical inference is the process by which we draw a conclusion about some measure of a *population*¹ based on a sample value. The measure might be a variable, such as the average or *mean* amount of money that consumers plan to spend on a new car, or an attribute, such as the *percent* of consumers favoring foreign cars. The purpose of sampling is to estimate these same characteristics for the population from which the sample is selected.

The population measure is the *parameter*, while the sample measure is called a *statistic*. We will first consider the problem of estimating the *arithmetic mean* of a population from the mean of a sample. This is called a *point estimate*, since it endeavors to provide the best single estimated value of the parameter. An *interval estimate*, on the other hand, proceeds by specifying a range of values. Thus, after testing a sample of steel rods, we may make a point estimate that the mean breaking strength of all such rods is 10 pounds; but we might also make an interval estimate that the mean for all rods probably lies in the interval from 8 to 12 pounds, as described later.

¹ "Population" and "universe" are usually considered synonymous. The newer term "population" will be used in this discussion. These terms refer here to inanimate objects as well as to living beings.

Sample information may be used for either of two purposes—*reporting* or *decision-making*. In a reporting role, the sample estimates (either point-estimates or interval-estimates) are presented for the information of others. Government statistics (e.g., on unemployment) are a good example of the use of sample data for this purpose. The sample information may be used also in this context to corroborate some point in exposition, as when a social scientist presents such information to help in drawing some conclusion. *Confidence intervals* are presented in this chapter for the purpose of reporting sample evidence and drawing conclusions therefrom.

On the other hand, the sample information may be incorporated directly in a decision-making procedure. *Tests of hypotheses* will be described in Chapter 12 as a means of decision-making, as well as reporting sample findings. Or, to go a step further, the sample may be combined with the prior judgments of the decision-maker and the economic consequences of various actions to arrive at the best decision. Chapters 15 and 16 incorporate samples in this decision-making context.

SAMPLING ERROR AND BIAS

A sample rarely produces, without error, the exact information needed for decision-making. Some reasons for the deviation of sample results from the true population values are as follows:

1. ***Sampling Error.*** Sampling error is the random or chance error that occurs when we take a sample rather than testing the whole population. A sample is only partially representative of the larger population from which it is taken. And any two samples will differ from each other, since they will contain different elements of the population.

If a probability sample (see below) is taken properly, sampling error can be controlled and measured. In general, sampling error can be reduced by increasing the size of the sample. Since larger samples are more costly, a key element of sample design is balancing the cost of a sample with the value of the information the sample provides.

2. ***Bias in the Manner in Which the Sample is Taken.*** If the sample is drawn in such a way that some elements of the population cannot be drawn at all, then some bias will arise. The classic example of this bias is the poll taken in 1936 by the *Literary Digest*. The *Literary Digest* mailed out 10 million cards and received about 2.3 million returns. On the basis of this sample, a victory by Alfred Landon for President was predicted. Actually, Roosevelt won with about 60 percent of the vote. The trouble with the *Literary Digest* sample was that it was

taken from lists of telephone subscribers and automobile registrants—in general, a higher income group not representative of the overall population of voters.

Sometimes in business research it is almost impossible to eliminate this kind of bias. Consider the firm that wishes to test a new advertising campaign. Very often it is economically feasible to select only one or two cities in which to test the new program. If the city selected is Atlanta, we obviously cannot measure the effect in Seattle. It is necessary to use business judgment to select an area that is “representative” of the nation as a whole. Experience in similar surveys and advertising programs would be useful in making this judgment.

3. Bias Due to Nonresponse. In almost any survey there are a number of items which are drawn in the sample for which no information is available. These may be people who do not mail back a questionnaire or who slam the door in the face of the interviewer. If these items are ignored, considerable bias may result, for the nonrespondents may be entirely different from the respondents. Thus, a significant part of the population may be ignored.

Every effort should be made to reduce nonresponse. This can be partly done in the design stage of the survey by careful wording and pretesting of questionnaires and instruction sheets to those conducting the survey. Training of survey personnel is also helpful in reducing nonresponse. And finally, extensive searches and callbacks should be employed.

4. Measurement Bias. Considerable bias can be introduced into a survey if the measurement instrument (questionnaire, interview, counting procedure, etc.) is not accurate, that is, does not measure what is intended. Consider the interviewer who found that most of those he interviewed said that they had never borrowed money from a loan company, despite the fact that the interviewer's list was drawn from a loan company's files. We must be equally careful in asking people how they will vote, whether they will buy our product, and so on.

Careful preparation of the questionnaire will help reduce this kind of bias, as described in Chapter 2. In addition, a pretest and a follow-up check on the measuring instrument and the results of the survey are essential.

Control of nonsampling bias from the three sources just noted is of crucial importance in survey work. It is much better to take a small sample which is relatively free of bias than to take a much larger sample with unknown bias. It is a common misconception to suppose that a large sample will iron out biases. It is not so!

SIMPLE RANDOM SAMPLING

There are many effective methods of selecting samples, and these may be used in various combinations. The sample may be selected from the population as a whole, or it may be selected from certain parts or *strata* of the population. In either case, the sample may be selected at random, according to somebody's judgment, or by other methods. The individuals selected may be drawn one at a time or in clusters, such as the residents of selected city blocks. The clusters may be enumerated completely, or they may be subsampled by selecting, say, the head of every third household in the block. Thus, these procedures provide a great variety of sample designs. One distinction is made between *probability samples* and others. A probability sample is taken in such a manner that elements of the population have a specific probability of being included in the sample. A measure of sampling error can be estimated for most probability sampling methods. Other methods rely on the judgment of the one selecting the sample or on other nonrandom procedures. While such samples may be quite useful, there is no accurate way of measuring their sampling error.

The basic concepts of statistical inference are applied to simple random samples in Chapters 11 to 13. While simple random sampling is not often used alone in business and economic research, it is important because it illustrates the fundamental principles of sampling and is a basic part of more complex types of sample design that are described in Chapter 14.

A *simple random sample* of n units is one selected from a population in such a way that each combination of n units has an *equal* chance of being selected. Thus, in selecting a simple random sample of five bolts from a shipment, every combination of five bolts in the shipment must have the same chance of selection. The bolts could not be picked only from certain boxes or just from the top of the pile.

This method is sometimes called "unrestricted" random sampling because units are selected from the population as a whole without any restriction, whereas procedures like stratification and clustering introduce restrictions (e.g., grouping the population before the sample is selected) designed to increase the precision of the sample or to reduce its cost.

Random sampling does not mean haphazard selection. Interviewing passers-by on a downtown street corner does not provide a random sample of a city's population because stay-at-homes have less chance of being interviewed than downtown shoppers or businessmen.

Random selection is determined objectively by some equivalent of a

game of chance. For example, the residents of a city block might be numbered from 1 to 72 and a roulette wheel could be spun ten times to determine the choice of ten persons to be interviewed. However, selections are usually made from a *table of random numbers*. Such a table is just as efficient as operating a game of chance and is more convenient. In constructing a table of random numbers, the digits from 0 to 9 are drawn by some randomizing device so that each number is independent of any other. The RAND Corporation, for instance, programmed an electronic computer so as to produce the random numbers listed in its book *A Million Random Digits*. Table 11-1 is a section of another such table. (See Appendix L at the end of this book for a larger table.)

Table 11-1

RANDOM NUMBERS

03	47	43	73	86	36	96	47	36	61	46	98	63	71	62
97	74	24	67	62	42	81	14	57	20	42	53	32	37	32
16	76	62	27	66	56	50	26	71	07	32	90	79	78	53
12	56	85	99	26	96	96	68	27	31	05	03	72	93	15
55	59	56	35	64	38	54	82	46	22	31	62	43	09	90
16	22	77	94	39	49	54	43	54	82	17	37	93	23	78
84	42	17	53	31	57	24	55	06	88	77	04	74	47	67
63	01	63	78	59	16	95	55	67	19	98	10	50	71	75
33	21	12	34	29	78	64	56	07	82	52	42	07	44	38
57	60	86	32	44	09	47	27	96	54	49	17	46	09	62
18	18	07	92	46	44	17	16	58	09	79	83	86	19	62
26	62	38	97	75	84	16	07	44	99	83	11	46	32	24
23	42	40	64	74	82	97	77	77	81	07	45	32	14	08
52	36	28	19	95	50	92	26	11	97	00	56	76	31	38
37	85	94	35	12	83	39	50	08	30	42	34	07	96	88
70	29	17	12	13	40	33	20	38	26	13	89	51	03	74
56	62	18	37	35	96	83	50	87	75	97	12	25	93	47
99	49	57	22	77	88	42	95	45	72	16	64	36	16	00
16	08	15	04	72	33	27	14	34	09	45	59	34	68	49
31	16	93	32	43	50	27	89	87	19	20	15	37	00	49

SOURCE: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (6th ed.; London: Oliver & Boyd, 1963), Table XXXIII, Random Numbers (1). This is part of a much larger table.

How to Use a Table of Random Numbers

To illustrate the use of this table, suppose you wish to select a random sample of 6 households from a city block of 78 households, as part of a market survey to determine brand preferences for frozen foods. First, list all households by address and number them from 01 through

78. Second, take a page from a table of random numbers, and choose a starting point at any arbitrary point²—say, the thirteenth column, fifth row, in Table 11-1. This number is 43. Third, go down this column and the next columns to the right (or go in any predetermined direction) until you have selected six numbers between 01 and 78, with no repetitions.

Beginning with 43, the next number down is 93, but it is ineligible, being larger than 78, so continue with 74, 50, 07, 46, 86 (ineligible—larger than 78), 46 (ineligible—already selected), and 32—a total of six eligible numbers. Thus, the numbers of the households to be surveyed are 7, 32, 43, 46, 50, and 74.

If there are exactly 100 items in the population, read "00" as 100. If there are more than 100 items, combine adjacent columns as necessary to form larger numbers. Thus, in the upper-left corner of Table 11-1, the columns beginning 034 could be used for three-digit numbers, or those beginning 0347 for four-digit numbers.

HOW SAMPLE MEANS ARE DISTRIBUTED

The use of the sample mean to make inferences about the population mean is a common problem in statistical inference. The following methods apply strictly to the *means* of simple random samples; they will be adapted to percents and to other types of samples in later chapters. Therefore, the term "sample mean" in this chapter will refer to the arithmetic mean of a simple random sample.

The following symbols will be used:

	<i>Sample</i>	<i>Population</i>
Arithmetic mean.....	\bar{X}	μ
Standard deviation.....	s	σ
Standard error of the mean.....	$s_{\bar{x}}$	$\sigma_{\bar{x}}$
Number of items.....	n	N

If we are interested in estimating *totals* for a population, we simply multiply the estimate of the mean and standard error of the mean by the number of items in the population. Thus:

	<i>Sample Estimate</i>	<i>Population Value</i>
Population total.....	$T = N\bar{X}$	$N\mu$
Standard error of population total.....	$s_T = Ns_{\bar{x}}$	$N\sigma_{\bar{x}}$

Inferences about a population are usually made from a *single* sample. This is only one of a large number of samples that might be drawn from

² Ideally, the starting point should be selected by a game of chance. In practice, however, an arbitrary choice is generally considered satisfactory.

the same population. By studying the variation of the means of all these samples, we can infer within what limits *our* sample mean is likely to fall. The means of all possible samples drawn from a given population may be grouped in a frequency distribution. This is called the *sampling distribution of the mean*. The mean and standard deviation of this distribution will describe the behavior of the sample means.

Table 11-2

SAMPLING THE DIAMETERS OF 565 BALL BEARINGS

DIAMETER*	NUMBER OF BALL BEARINGS IN—						
	Popula- tion	1st Sample	2d Sample	3d Sample	4th Sample	5th Sample	5 Samples Combined
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
-6	1	1	...	1	2
-5	4	1	...	2	3
-4	15	...	2	1	1	...	4
-3	38	2	1	1	4	3	11
-2	70	8	7	5	3	10	33
-1	97	9	7	12	7	11	46
0	115	12	11	11	10	6	50
1	97	9	11	10	8	7	45
2	70	5	4	6	9	4	28
3	38	1	5	1	4	4	15
4	15	4	2	...	3	2	11
5	4	1	...	1
6	1	1	1
Number of ball bearings	565	50	50	50	50	50	250
Average diameter*	0	+.14	+.20	-.18	+.52	-.42	+.05

* Difference from specification (0.250 inches) in thousandths of an inch.

An Experiment

To illustrate the sampling distribution of the mean when the population is known, consider the following experiment:

A manufacturer of electrical equipment receives shipments of ball bearings from a steel company for use in electric fans. Specifications call for these balls to average one-quarter inch in diameter, and none of them must deviate from the specification by more than a given toler-

ance. Since it is not feasible to measure every ball bearing, it is necessary to depend on sample inspection to avoid acceptance of unsatisfactory shipments.

The inspection supervisor wished to illustrate the sampling principles involved as part of the training program for inspectors. Accordingly, he selected one shipment of 565 ball bearings as the population. He then had the whole lot measured with automatic calipers. The results are shown in Table 11-2, columns 1 and 2. Thus, only one of the 565 balls was six thousandths of an inch below specification, four balls were five thousandths below, and so on; the average of all the balls (last row) was exactly equal to the specification.

Samples of 50 steel balls each were then selected at random from the bin containing the shipment, and their diameters were measured. After each 50 were selected, they were returned to the bin and thoroughly mixed so that the next sample could be selected from the same population as the first sample. In all, 100 samples of 50 balls each were selected.

The results of the first 5 of the 100 samples are shown in columns 3 to 7 of Table 11-2. Each of these samples differs from the others, and none of them is a perfect replica of the population. The mean diameter for each sample is shown in the last row.

The Three Distributions. It is important to distinguish the three different distributions illustrated by this experiment. They are shown in Chart 11-1. First is the distribution of ball-bearing diameters (X) in the population itself—curve A. The figures are taken from Table 11-2, column 2. Frequencies are plotted as percentages of the total, on the Y axis, for comparability with curve B. (The curve would have been smooth if the ball bearings had been measured exactly rather than to the nearest 0.001 inch.) This population is normal, with its mean μ equal to zero. Other populations may be skewed or otherwise irregular.

Second is the distribution of the X values in a sample drawn from this population, such as the fourth sample in Table 11-2, shown in curve B. The sample distribution has somewhat the same general shape as the population, but it is more irregular, and its mean (\bar{X}) differs from the true mean (μ) because of sampling errors. As the sample size increases (e.g., Table 11-2, column 8), the shape of the sample distribution approaches more and more closely that of the population distribution, whether the latter be skewed or what not. Both the mean and the standard deviation of the sample also approach the population values.

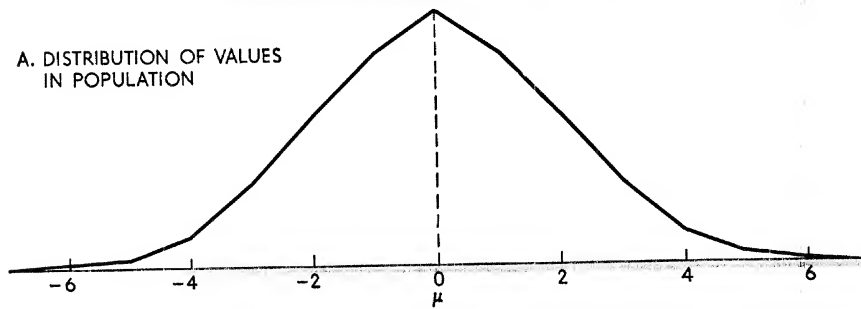
Third is the sampling distribution of the *means* (\bar{X}) of a great many

samples (curve C) of size $n = 50$ that can be drawn from this population. This curve shows the distribution of 100 sample means. It has been drawn with a smaller area than that under the other curves; otherwise it would be awkwardly tall. The five sample means shown in the bottom row of Table 11-2 fall well within the range of curve C. The mean of this distribution is very close to that of the population, and its dispersion

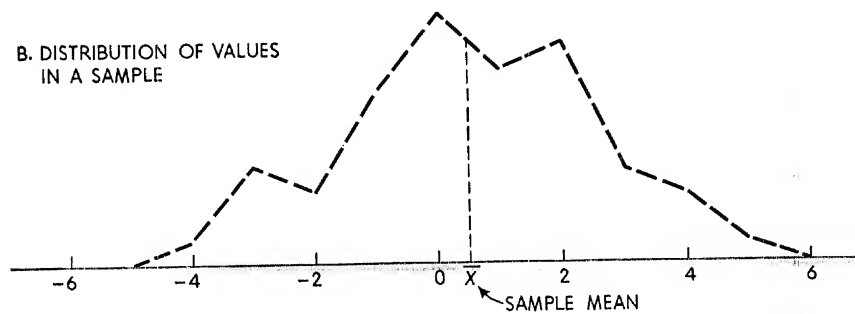
Chart 11-1

THE THREE DISTRIBUTIONS INVOLVED IN ESTIMATING
THE MEAN
BALL-BEARING DIAMETERS (TRUE MEAN = 0)

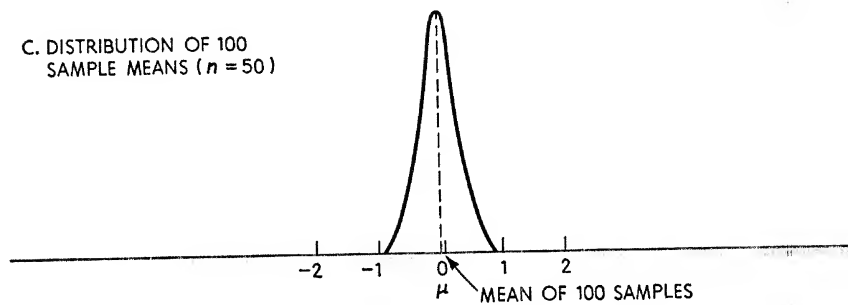
A. DISTRIBUTION OF VALUES
IN POPULATION



B. DISTRIBUTION OF VALUES
IN A SAMPLE



C. DISTRIBUTION OF 100
SAMPLE MEANS ($n = 50$)



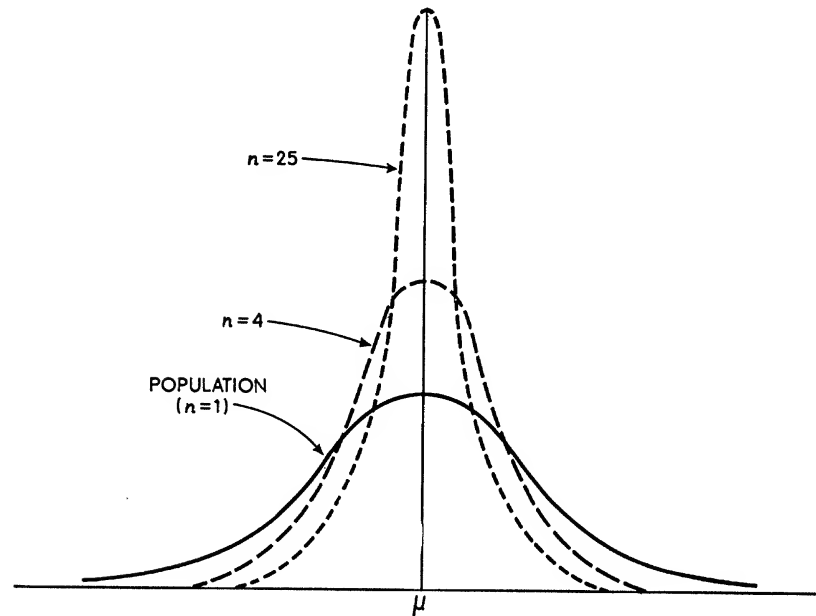
Unit: Thousandths of an inch differences from specification.
Source: Table 11-2 and related data.

or standard deviation is much less than that of curve A or B. If all possible samples of size 50 were drawn from this population, the distribution shown in curve C would be smoother, and nearly normal.

As the sample size increases, the distribution of sample means becomes still narrower in spread, and more normal in shape, as described below. Chart 11-2 shows how the sample means from a normal population tend to cluster more closely about the population mean as the sample size increases. The three curves in Chart 11-2 have the same

Chart 11-2

SAMPLING DISTRIBUTIONS OF THE MEANS OF SAMPLES
OF SIZE $n = 4$ AND $n = 25$, COMPARED WITH
DISTRIBUTION OF A NORMAL POPULATION



area and are all normal, but they differ markedly in dispersion.

Sampling Concepts. The ball-bearing experiment illustrates several concepts in sampling:

1. Each of the means is approximately, but not exactly, equal to the population mean. Of the 100 samples selected in the larger study (not reported here in detail), only 5 exactly equaled the population in mean diameter, while 53 were above and 42 were below.
2. The sample means cluster much more closely about the population mean than do the original values. Thus, the means in the last row of the table vary only from -0.42 to $+0.52$, while the individual

diameters (columns 1 and 2) range from -6.0 to $+6.0$. Hence, the standard deviation of the sample means is much smaller than the standard deviation of the original values.

3. If larger samples were taken, their means would cluster still more closely around the population mean since the positive and negative errors of sampling tend to offset each other. This is illustrated by combining the five samples shown to obtain the larger sample of 250 balls listed in column 8. The mean of this larger sample is $+0.05$, a result which is much closer to the population value (0) than is any of the means of the 5 samples of 50. The overall average of the 100 sample means proved to be $+0.02$, which is closer yet to the population mean.

Thus, *the larger the sample, the closer its mean is likely to be to the mean of the entire population*, and the greater the precision of the sample mean. It can be shown that if all possible samples of a given size are drawn from a population, the arithmetic mean of the sample means will equal the population mean.

4. The distribution of sample means follows a normal curve. More precisely, if a number of random samples of size n are drawn from a given population, their means tend to form a normal distribution, provided (1) the size of sample is large³ and (2) the population is not unduly skewed. If the population is skewed, the distribution of sample means will be much less skewed, in inverse proportion to the size of the sample. Thus, for samples of size 50 the distribution of means will only be $\frac{1}{50}$ as skewed as the population (i.e., $n = 1$).⁴

The arithmetic mean therefore tends to be normally distributed as n increases in size, almost regardless of the shape of the original population. This principle is called the *central limit theorem*. It applies to the distribution of most other statistics as well, such as the median and standard deviation (but not the range). The central limit theorem gives the normal distribution its central place in the theory of sampling, since many important problems can be solved by this single pattern of sampling variability.

The distribution of sample means being normal, or nearly so, it can be completely described by its mean and its standard deviation. Furthermore, these values may be estimated from a single large random sample, as described under "The Standard Error of the Mean" below.

³ In many cases a size of 30 is large enough, but no exact number can be given; it depends in part on the population distribution.

⁴ See F. E. Croxton and D. J. Cowden, *Applied General Statistics* (2d ed.; New York: Prentice-Hall, 1955), p. 627.

The Sample Mean as an Estimator of the True Mean

When we select a statistic such as the mean to estimate the population value, we ordinarily expect it to satisfy two criteria:

1. The statistic should, on the average, give the "correct" answer—the population value. That is, the mean of a distribution of all possible means for a given size of sample—that is, the *expected value*—should equal the population value. Such an estimate is said to be *unbiased*. Means of random samples are unbiased estimators of the true means. Thus, in Table 11-2, the expected value is the overall mean of all possible sample means, each representing 50 ball bearings. This is zero, the same as the population mean. The mean of an individual sample, then, whatever its value, is said to be an unbiased estimator.

2. The second criterion states that the sampling distribution of the statistic be concentrated as closely as possible about the true population value. Such a statistic is said to be *efficient*. It can be shown that the sample mean is a more efficient estimator of the parameter than the sample median in a normal population, since the sample values cluster more closely about the population value. In Chart 11-1, panel C, a distribution of sample medians would have a wider spread than that shown for the means.⁵ (The median may be more efficient, however, for sharply peaked, long-tailed distributions, as noted in Chapter 5.)

THE STANDARD ERROR OF THE MEAN

The standard deviation of the distribution of sample means is called the *standard error of the mean*. (The word "error" is used here in place of "deviation" to emphasize that variation among sample means is due to sampling errors.) The standard error measures (inversely) the *precision* of the sample estimate, that is, how closely the sample value is likely to approach the true value.⁶ The smaller the standard error, the greater the precision. Where the population is large in relation to the sample size, the formula for the standard error of the mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

⁵ The standard error of the median is 1.25 times that of the mean in a normal population.

⁶ "Precision" or "reliability" as used in statistics means how closely we can reproduce from a sample the results that would be obtained if we took a complete census, using the same methods of measurement, interview procedures, etc. The "accuracy" of a survey takes into account these sampling errors as well as nonsampling errors arising from bias due to methods of measurement, questionnaire design, etc. that would affect the census as well as the sample. We can only measure precision, but it is the overall accuracy that we attempt to maximize in designing surveys.

where σ is the standard deviation of X in the population and n is the size of the sample.

Thus, in the ball-bearing example the standard deviation of the population (Table 11-2, column 2) is

$$\sigma = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{\frac{2,190}{565}} = 1.969 \quad (\text{unit} = 0.001'')$$

Then, for samples of size 50, the standard error of the mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.969}{\sqrt{50}} = 0.278$$

and for samples of size 250,

$$\sigma_{\bar{x}} = \frac{1.969}{\sqrt{250}} = 0.124$$

The standard error of the sample means, therefore, varies directly with the standard deviation of the population σ , and inversely with \sqrt{n} . By increasing the sample size, the standard error of the mean can be reduced to any desired level. However, the reduction is not pro rata. The sample size must be quadrupled to cut the standard error in half.

Finding the Standard Error of the Mean When σ Is Unknown

In practice, the standard deviation of the population (σ) is usually unknown, but it can be estimated as being equal to the standard deviation of a single large sample (s). That is, instead of $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, we can say

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where $s_{\bar{x}}$ is the standard error of the mean estimated from a single sample, and s is the standard deviation of the sample.⁷

Thus, for the first sample in Table 11-2, .

$$s = \sqrt{\frac{\sum fx^2}{n-1}} = \sqrt{\frac{161}{49}} = 1.81$$

⁷ Sometimes n is used instead of $n-1$ in the formula for s , e.g., $s = \sqrt{\sum fx^2/n}$. In this case use $s_{\bar{x}} = s/\sqrt{n-1}$ to achieve the same result as above. That is, by combining the two formulas, $s_{\bar{x}} = \sqrt{\sum fx^2/n(n-1)}$ in either case. (Omit f in formulas for ungrouped data.)

and

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1.81}{\sqrt{50}} = 0.256$$

This estimate of the standard error of the mean differs by 8 percent from the true $\sigma_{\bar{x}}$ of 0.278.

Again, for the combined sample of 250,

$$s = \sqrt{\frac{1,017}{249}} = 2.021$$

and

$$s_{\bar{x}} = \frac{2.021}{\sqrt{250}} = 0.127$$

For the larger sample, the estimated standard error of the mean differs by only 2 percent from the true $\sigma_{\bar{x}}$ of 0.124. This illustrates the principle that the standard error of the mean can usually be estimated satisfactorily from the standard deviation of a single large sample (the larger the better) when the standard deviation of the population is unknown.

Effect of Population Size. The above formulas for $\sigma_{\bar{x}}$ and $s_{\bar{x}}$ are correct if the population is infinitely large, or if the sampling is carried out with replacement, which amounts to the same thing. Sampling with replacement means that after an item is selected it is replaced and has a chance of being selected again. These formulas are also substantially correct when the sample is a small percent—say less than 5 percent—of a finite population. Thus far, the ball-bearing experiment has been treated as if its population were infinite.

Where the sample comprises a large proportion of the population and is done without replacement, the expression σ/\sqrt{n} should be multiplied by $\sqrt{(N-n)/(N-1)}$, or approximately $\sqrt{1-n/N}$, where n is the sample size and N is the population size. That is,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} \quad \text{for finite populations.}$$

The term $1 - n/N$ is the proportion of the population not included in the sample, and is called the finite population correction.⁸ Its use always reduces the standard error.

⁸ See M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory* (New York: John Wiley, 1953), Vol. I, pp. 122–24; and W. A. Wallis and H. V. Roberts, *Statistics, A New Approach* (New York: The Free Press, 1956), pp. 368–71. The finite population correction is also called the finite population factor, finite multiplier, and finite sampling correction.

For example, since each sample of 50 ball bearings in Table 11-2, columns 3 to 7, was drawn without replacement from the population of 565 balls, we should have

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{1.969}{\sqrt{50}} \sqrt{1 - \frac{50}{565}} \\ &= 0.278 \times 0.955 \\ &= 0.265 \cdot\end{aligned}$$

instead of 0.278 for sampling with replacement.

Thus, the precision of the sample estimate, measured by $\sigma_{\bar{x}}$, is determined not only by the absolute size of the sample but also to some extent by the proportion of the population sampled. This is in accordance with common sense. A 10 percent sample certainly seems more reliable than a 5 percent sample.

In most actual surveys, however, the sample is such a small percent of the population that n/N is negligible, and $\sigma_{\bar{x}}$ is virtually equal to σ/\sqrt{n} . Hence, the reliability of a sample usually depends almost entirely on the absolute size of the sample and *not* on the percentage of the population sampled. In planning a market survey of consumers in a large city, one should ask questions like "Is a sample of 1,000 big enough?" and not "Is a 10 percent sample big enough?" The size of the city makes little difference.

How $\sigma_{\bar{x}}$ Is Used

The standard error of the mean in the ball-bearing example is 0.265 thousandths of an inch for samples of size 50. Since 0.265 is the standard deviation of all possible means of size 50, and the distribution of means of large samples is normal, we can say what proportion of the sample means lies within any given interval of the true (population) mean. In this case the true mean is known ($\mu = 0$). Then 68.27 percent of the sample means fall within one standard error ($\sigma_{\bar{x}}$) of the true mean, that is, from $+0.265$ to -0.265 . As noted in Chapter 8, this means that there is a *probability* of about 68 percent—or 68 chances out of 100—that a *single* sample mean will fall within the interval of $\mu \pm \sigma_{\bar{x}}$, or 0 ± 0.265 ; and so on for any other degree of probability desired.

These figures also show just how much more closely the sample means cluster than do the individual ball-bearing diameters. While 68 percent of the means lie within $\sigma_{\bar{x}}$ or 0.265 thousandths of an inch from the true mean, the same percentage of individual ball bearings lie within σ or 1.969 thousandths of the true mean—a far wider spread.

If the distribution of the population is not normal, the above figures are still approximately correct for larger samples. In an experiment at the University of California, Berkeley, some 3,000 independent samples of 30 items each were drawn at random (using a table of random numbers) from a skewed population consisting of 200 weekly earnings figures for a group of wage earners and clerical workers in the San Francisco Bay Area. The population values ranged from \$17.50 to \$116.91 a week and averaged \$57.95. The arithmetic mean, the standard deviation, and the approximate standard error of the mean $s_{\bar{x}}$ were computed for each sample. The question then arose: What percentage of the 3,000 sample means fell within various multiples of the standard error around the true population mean μ of \$57.95? The results were as follows:

	$\mu \pm s_{\bar{x}}$	$\mu \pm 2s_{\bar{x}}$	$\mu \pm 3s_{\bar{x}}$
Theoretical expectancy.....	68.27%	95.45%	99.73%
Experimental results.....	68.4 %	95.2 %	99.6 %

This shows a remarkable agreement between fact and theory, despite the fact that (1) the sample size was but 30 items; (2) the sample standard deviation s was used, instead of the true population value σ ; and (3) the population was not normally distributed. The theory therefore works well in practice. For smaller samples, however—say when n is under 30—the above values may have to be adjusted, as is described in Chapter 13.

The corresponding results for any other probability or interval in the sampling distribution of means can be found in Appendix D, just as we previously did for individual values. For example, within what interval will exactly 95 percent of the sample means fall in the ball-bearing case ($n = 50$)? Since the proportion 0.95 lies on both sides of the population mean, look up half this amount, 0.475, for the proportion on one side of the mean, in the body of Appendix D. The interval is then $\pm 1.96\sigma_{\bar{x}}$ or ± 0.519 thousandths of an inch.

It is customary to state probabilities in such round numbers as 95 or 99 percent, so the following relationships in a normal distribution are important:

Mean $\pm 1.96\sigma$ includes 95.0 percent of the area
Mean $\pm 2.58\sigma$ includes 99.0 percent of the area

These are often used instead of the statements that the mean $\pm 2\sigma$ includes 95.45 percent of the area and $\pm 3\sigma$ includes 99.73 percent.

When the population mean is *not* known, and we use a sample mean

to estimate it, we can only say that 68 percent of the sample means lie within one standard error of the true mean, wherever that may be, and similarly for other intervals. Nevertheless, we will see in the next section how this information about the spread of sample means around the unspecified true mean can be used to make satisfactory estimates of the true mean.

CONFIDENCE INTERVALS

It is often necessary to estimate the unknown mean (or other parameter) of a population. To do so, we need both the sample value and a measure of the margin of error to which this value is subject. This may be done as follows:

1. Find the mean \bar{X} and its standard error ($s_{\bar{X}} = s/\sqrt{n}$) from a large random sample as point estimates of the population values.
2. Specify a zone based on \bar{X} and $s_{\bar{X}}$ within which we may be confident that the true population mean does lie. This is called a *confidence interval*. The end points of this interval are called *confidence limits*.
3. State the probability—say, 95 or 99 percent—that such a zone will include the population mean. This probability is called the *confidence coefficient* or *level of confidence*. It must be set in advance. Each confidence interval that may be chosen has an associated probability of including the population mean—the wider the interval the greater the probability. Thus, the zone $\bar{X} \pm 1.96 \sigma_{\bar{X}}$ is the “95 percent confidence interval.” This relationship is based on the fact that 95 percent of all sample means tend to fall within $1.96 \sigma_{\bar{X}}$ of the population mean, where $\sigma_{\bar{X}}$ is the true standard error of the mean. Similarly, the zone $\bar{X} \pm 2.58 \sigma_{\bar{X}}$ is the “99 percent confidence interval.” The zone for any other confidence coefficient may be found in Appendix D. The selection of the appropriate confidence coefficient is discussed on page 267.

For example, we wish to estimate the mean diameter of the population of ball bearings in Table 11-2, which is assumed to be unknown. We take sample No. 1 (column 3) and proceed as above. (All units are in thousandths of an inch.)

$$\bar{X} = +0.14$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} = \frac{1.81}{\sqrt{50}} \sqrt{1 - \frac{50}{565}} = \frac{1.81}{7.07} (0.955) = 0.244$$

Use this value as an estimate of the true standard error of the mean $\sigma_{\bar{X}}$. The error involved is a minor one for larger samples.

Compute $\bar{X} \pm 1.96 s_{\bar{X}}$ as the 95 percent confidence interval for the population mean:

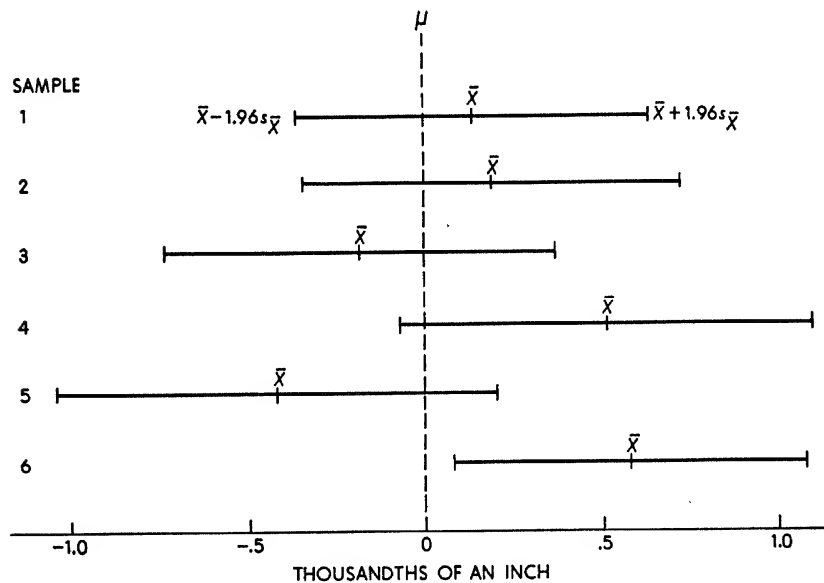
$$\bar{X} + 1.96 s_{\bar{X}} = 0.14 + 1.96(0.244) = 0.14 + 0.48 = +0.62$$

$$\bar{X} - 1.96 s_{\bar{X}} = 0.14 - 1.96(0.244) = 0.14 - 0.48 = -0.34$$

Our best point estimate of the population mean is therefore the sample mean, $+0.14$, but this estimate is subject to a margin of error defined by the 95 percent confidence limits of $+0.62$ and -0.34 . This probability statement needs some interpretation. For any particular sample, the confidence interval either includes the population mean or it does not—we do not know. The *objective* probability is either 100 percent or zero (in this case it does, since we know the population mean is zero). Strictly speaking, the statement means that if a very large number of samples of size n are drawn, and the confidence interval is computed for each, 95 percent of these intervals will include the population mean.

Chart 11-3

95 PERCENT CONFIDENCE LIMITS FOR THE POPULATION
MEAN OBTAINED FROM 6 SAMPLE MEANS OF
BALL-BEARING DIAMETERS ($n = 50$)



Source: Table 11-2 (except sample 6).

On the other hand, using a *subjective* interpretation of probability, we can make the more straightforward statement that there is a 95 percent chance that the population mean lies within the confidence interval. In other words, one should be willing to bet, with odds of 19 to 1, that the population mean lies in the interval $+0.62$ to -0.34 based only on the sample information.

Chart 11-3 shows the means and confidence limits for this sample and for the other four samples of 50 ball bearings listed in Table 11-2.

The means and intervals all vary, but the latter all include the population mean μ , shown as a dashed line. The confidence interval for a sixth sample, however (not shown in Table 11-2), fails to include the true mean. Of all such possible confidence intervals, then, 95 percent include the population mean.

The confidence interval around a sample mean might be likened to a quoit aimed at a peg—the population mean. Then 95 percent of the quoits will ring the peg. If a bigger quoit is used—say the wider 99 percent confidence interval of $\bar{X} \pm 2.58s_{\bar{X}}$ —then 99 percent of the quoits will be ringers.

A 99 percent confidence interval can be computed as $\bar{X} \pm 2.58 s_{\bar{X}}$ and similarly for any other confidence coefficient, using the table of areas under the normal curve. The 99 percent interval for ball-bearing sample No. 1 is

$$\bar{X} \pm 2.58s_{\bar{X}} = +0.14 \pm 2.58(0.244) = +0.14 \pm 0.63$$

Hence, we can say, in subjective terms, that there is a 99 percent chance that the population mean lies between the confidence limits of -0.49 and $+0.77$.

Which Confidence Coefficient Should Be Selected?

Raising the confidence coefficient from 95 to 99 percent increases our degree of assurance that the confidence interval contains the population value, but it also makes our estimate less precise, since the confidence interval itself has been widened by 32 percent (i.e., from 1.96 to 2.58 standard errors). In deciding which confidence level to use, we must understand that the primary purpose of the confidence interval is to report or communicate to others the results of the sample. The confidence interval is a convenient way of expressing the sampling error by giving an interval that is likely to include the population mean. The confidence level chosen therefore is sometimes rather arbitrary. In particular, the 95 percent level is often used in the social sciences, and the 99 percent level in the natural sciences where precision is higher. Other

levels should be chosen, however, when we can balance the value of a precise estimate against the cost of missing the true value.

Any economic or business report that cites the mean (or other statistic) of a probability sample should give the reliability of this value in terms of a confidence interval or some other use of $\sigma_{\bar{x}}$ as a measure of the sampling error. For example, a Census Bureau's *Monthly Report on the Labor Force* says, "The chances are about 19 out of 20 that the difference between the estimate and the figure which would have been obtained from a complete census is less than the sampling variability indicated below" (followed by a table showing various sample sizes and the corresponding 95 percent confidence intervals). A statistic having a large sampling error may be useless; at any rate, the error should be stated. The report should also point out that this reliability measure does not include the effect of bias due to nonsampling errors in sample design, incomplete coverage of sample, bias of respondent, etc. These errors should be discussed in qualitative terms.

Errors in Confidence Intervals. The confidence intervals just described may be inaccurate because (1) the standard error of the mean estimated from a single sample is not equal to the true standard error and (2) the sample means may not be quite normally distributed. These errors are appreciable in small samples, but they become insignificant in larger samples. Thus, in the example cited above, increasing the sample size from 50 to 250 reduced the discrepancy in the standard error of the mean from 8 to 2 percent.

HOW BIG SHOULD A SAMPLE BE?

In planning a sample survey, is it necessary to sample 100 items? 1,000? Or all we can afford? The answer depends mainly on two factors: (1) the economic value of the information contained in the sample and (2) the cost of sampling. The value of sample information and the cost of the sample both increase as sample size increases. The optimum sample size is that which balances the cost and value of the sample. Determination of optimum sample size is discussed in Chapter 16. In this section we will discuss two related questions: (1) How large a sample is needed to obtain a given degree of precision in the sample estimate? and (2) How to balance sample precision against the cost of sampling?

The relation between precision of the sample mean and size of sample is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

To estimate how big n should be, there are three steps:

1. Determine how small the standard error of the mean $\sigma_{\bar{x}}$ must be in order to obtain the necessary precision. The precision depends on how the results are to be used.
2. Take a random sample of any convenient size and compute the sample standard deviation s as an estimate of σ , the population standard deviation.
3. Substitute the desired value of $\sigma_{\bar{x}}$ and the estimated σ in the above equation and solve for n . This size of sample will give the necessary precision. If a larger sample is then taken, its standard deviation can be used to provide a revised estimate of σ and hence $\sigma_{\bar{x}}$.

The size of the population is usually a negligible factor, as pointed out earlier. However, if the sample makes up more than 5 or 10 percent of the population, the finite population correction should be applied to the above equation.

As an example, suppose it is desired to estimate the population mean of ball-bearing diameters within 0.3 thousandths of an inch at the 99 percent confidence level (i.e., $2.58 \sigma_{\bar{x}} = 0.3$ thousandths). Take a sample of convenient size and compute s as an estimate of σ , for example, sample No. 1 in Table 11-2 where $n = 50$ and $s = 1.81$.

First, determine the desired $\sigma_{\bar{x}}$:

$$2.58\sigma_{\bar{x}} = 0.3 \quad \text{or} \quad \sigma_{\bar{x}} = \frac{0.3}{2.58} = 0.116$$

Now, substitute these values in the equation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and solve for n :

$$0.116 = \frac{1.81}{\sqrt{n}}$$

$$\text{Transposing,} \quad \sqrt{n} = \frac{1.81}{0.116} = 15.6$$

$$\text{Squaring both sides,} \quad n = 244$$

Therefore, a sample of 244 ball bearings (including the original 50) should be taken. Actually, in this example somewhat less than 244 would suffice since 244 is a significant part of the total population of 565 ball bearings and the finite population correction should be applied. In general, however, when we are sampling from large populations the finite population correction can be ignored.

The cost of a survey includes a constant factor—for setting up the project, overhead, etc.—and a variable factor (so much per item sampled). Suppose it costs \$300 to set up the ball-bearing inspection and \$1.00 per measurement. Then the total cost (C) in dollars is

$$C = 300 + 1n$$

The executive can then compare the cost with the precision of the sample result for various possible sizes of sample, in order to choose among them. Thus, for the ball-bearing example:

n	$s_{\bar{x}}^*$	Cost
50	0.256	\$350
250	0.127	550

*In thousandths of an inch.

Since the cost increases directly with the size of sample, and reliability increases only with the square root of sample size, there are diminishing returns, and at some point the slight increase in reliability will not justify the added cost of sampling.

Consumer surveys conducted by personal interview may cost many dollars per schedule, but where important decisions are at stake, the necessary precision may justify a costly survey. As a case in point, the Elgin National Watch Company suffered from foreign competition in the late 1950's and lost over \$8 million in 1957–1958. The company then spent \$50,000 for market surveys. According to *Time* (May 2, 1960):

The surveys showed that Elgin simply was not making what buyers wanted. Men were found to prefer round watches (most of Elgin's were rectangular), to like functional stainless steel, water- and shockproof cases (Elgin's were mostly yellow gold), to want sweep second hands (only 15% of Elgin's had them). . . .

The surveys also showed that consumers wanted cheaper watches. The company introduced new, competitively priced models, and in the year ended March 1, 1960, made net profits of \$815,000. Obviously, if a business decision as important as revising a product line depends on the results of market surveys, high precision is required and high cost is justified.

The reliability and cost of a survey depend not only on the size of sample but also on the sampling plan itself. The principal plans are discussed in Chapter 14. For example, instead of a simple random sample, the reliability of a given-sized sample can be increased by stratification, or the unit cost can be reduced by cluster sampling.

SUMMARY

Statistical inference is the process of making a generalization or prediction about a population value, or *parameter*, based on a sample value, or *statistic*. This may be a single-valued *point estimate*, or a range of values designated as an *interval estimate*. The process is first described for the mean of a simple random sample.

If all possible means of large samples are drawn from a population, the sampling distribution tends to follow a *normal curve*. The proportion of items that fall within a given area under the normal curve may be determined from Appendix D. This proportion represents *relative frequencies*, or the *probability* that a single item (e.g., a sample mean) will fall within the segment.

An experiment is presented to show how sample means cluster about the population mean—the cluster being closer and hence the precision greater for larger samples. The sampling distribution of the mean must be clearly distinguished from the distribution of individual values in the population or the somewhat similar distribution of individual values in the sample itself (Chart 11-1). The tendency of the sampling distribution of the mean to form a normal curve as n increases in size, whatever the type of population, is called the *central limit theorem*.

The sample mean is said to be an *unbiased* estimator of the population mean because its *expected* value equals the population value. The expected value is the mean of a distribution of all possible means for a given size of sample. The sample mean is also said to be *efficient* because its sampling distribution usually clusters more closely about the population value than does, say, the median.

The *standard error of the mean* (i.e., the standard deviation of all possible sample means) measures the precision of the sample estimate. It is related to the population standard deviation and the sample size as follows: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. However, since σ is usually unknown, the standard error of the mean can be estimated from the standard deviation of a single large sample by the formula $s_{\bar{x}} = s/\sqrt{n}$. This expression should be multiplied by $\sqrt{1 - n/N}$, the "finite population correction," if the sample size n is more than about 5 percent of the population size N .

Since sample means are normally distributed, the *probability* is 68 percent that a single sample mean will fall within the interval $\mu \pm \sigma_{\bar{x}}$. The probability for any other intervals can be found in Appendix D.

We can estimate that the population mean falls within a certain

confidence interval, based on the sample mean and standard deviation, with a predetermined probability—say, 95 or 99 percent—of being correct. Thus, $\bar{X} \pm 1.96 \sigma_{\bar{X}}$ is the 95 percent confidence interval for the mean—that is, if we state that the population mean falls within this zone, we will have a 95 percent chance of being correct. We can increase the confidence coefficient—say to 99 percent—but only at the cost of making the estimate less precise by widening the confidence interval. The choice depends on the problem. In any case, the confidence interval and coefficient should be stated in reporting the results of sample surveys.

The size of a sample can be determined by solving the equation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ for n , where $\sigma_{\bar{X}}$ measures the required precision, and σ is estimated from a trial sample. Since precision increases with \sqrt{n} and the cost of sampling increases with n , the precision and cost should be contrasted for several sizes of samples, as an aid in determining sample size. The question of optimal sample size is discussed further in Chapter 16.

PROBLEMS

1. Explain the following concepts:
 - a) Point estimate of the mean.
 - b) Sampling distribution of the mean.
 - c) Central limit theorem.
 - d) Standard error of the mean.
 - e) Confidence interval for the mean.
 2.
 - a) A machine, when in adjustment, produces parts that are normally distributed and have a mean diameter of 0.300 inches with a standard deviation of 0.040 inches. If the machine is in adjustment, what is the probability that the *mean* value of a random sample of 4 parts will fall between 0.290 and 0.304 inches?
 - b) What would happen to the standard error of the mean if we increased the sample size from 4 to 16?
 3. A random sample of 144 building bricks has a mean weight of 7.1 pounds and a standard deviation of 0.30 pounds. Is it likely that this sample comes from a brickyard that produces bricks with a mean weight of 7 pounds?
 4. "A sample of 40 from a population of 400,000 will give nearly as precise an estimate of the population mean as a sample of 40 from a population of 4,000, provided the standard deviations of the populations are the same." Is this statement reasonable? Give figures to support your answer.
 5. A random sample of 64 is drawn from the records of daily output of a large group of employees in order to estimate the population mean. The sample
-

shows a mean of 136 units and a standard deviation of 24 units. Calculate a 98 percent confidence interval for the mean output of all employees.

6. A random sample of 400 accounts receivable is selected from the 2,000 accounts due a firm. The sample mean is found to be \$165.50, with standard deviation of \$26.00. Set up a 95 percent confidence interval as an estimate of the population mean. Interpret the meaning of this interval.
 7. A survey is planned to determine the average annual family expenditures for medical expenses of employees in a given company within \$50, at the 90 percent confidence level. A pilot study provides an estimate of \$334 as the standard deviation of medical expenditures. How large a random sample is needed to yield an estimate with the necessary precision?
 8. The controller of a department store takes a sample of 64 monthly statements to be mailed to credit-card holders, and finds that the average amount owed is \$28, with standard deviation of \$12. How many accounts should he sample, in total, if he wishes to estimate the mean amount owed within a dollar, with only 1 chance in 20 of being outside that range?
 9. A certain company employs 400 executives. A sample of 36 is taken in order to estimate the average age of all the executives. The results of the sample are $\bar{X} = 51.0$ and $s = 4.0$ years. Calculate a 99 percent confidence interval for the mean age of all executives.
 10. A random sample of 225 orders from a batch received by a certain firm has an average size of \$12.74 and a standard deviation of \$2.45. Construct a 95 percent confidence interval for the average size of *all* orders received in this batch. (There were 625 orders in the batch.)
 11. How large a sample would be needed to estimate the mean life of a new type of incandescent lamp within 24 hours, with no greater risk than 1 chance in 20 of being wrong. The standard deviation of burning life is estimated at 200 hours.
 12. a) The planning commission in a city wished to estimate the mean number of inhabitants per dwelling unit in the city. It selected a simple random sample of 500 dwelling units and obtained the following results: $n = 500$, $\Sigma X = 2,200$, $\Sigma X^2 = 11,680$. Calculate a 95 percent confidence interval for the mean number of inhabitants per dwelling unit in the city.
b) Suppose that there were 10,000 dwelling units in the city. Set up a 95 percent confidence interval for the total population of the city.
(Hint: A population *total* can be estimated as $N\bar{X}$ and the standard error of this estimate as $Ns_{\bar{X}}$.)
 13. A random sample of 81 out of the 225 graduating seniors of a college received an average starting salary of \$620 a month, with a standard
-

deviation of \$80. Give a 90 percent confidence interval for the mean starting salary for all 225 graduating seniors.

14. Past experience indicates that the standard deviation of the amount of gasoline consumed per year by motorists in a certain area is 50 gallons. How large a sample must be taken for the estimate of the true mean consumption to have a 0.99 probability of being within 10 gallons of the actual true mean?
 15. The market research department of a certain company was allocated \$40,000 to make a survey on the potential sales of a new product. A sample of stores through which the company distributed its product was to be selected. The new product was to be introduced in this sample of stores and the sales noted over a period of three months. The average sales per store per month would then be used to estimate the total sales potential of the new product.
Suppose that it costs \$10,000, plus \$300 per store to conduct the sample. From past experience with similar products it is estimated that the standard deviation of sales per store per month is 68 packages of the product.
 - a) How large a sample can be taken for the amount allocated? What sampling error in the estimate of average sales per store per month can be expected?
 - b) Suppose that actually a sample of 80 stores was selected. In these stores, the average sales per store per month was 84 packages and the standard deviation of the monthly sales for the stores was 52 packages. Using these estimates, make an estimate of the total *annual* sales of this product if it were to be distributed through 80,000 stores. Calculate a 95 percent confidence interval about this estimate. (See hint given in problem 12 above.)
 - c) What probability would you assign to the possibility that estimated total annual sales was off by more than 8 million packages? By more than 5 million packages?
 16. A population is known to have a mean $\mu = 85$ and a standard deviation $\sigma = 15$.
 - a) What is the probability that the mean of a sample of size 25 will fall in the interval 83 to 87?
 - b) What is the probability that the mean of a sample of size 36 will fall in the interval 83 to 87?
 - c) What is the probability that the mean of a sample of size 81 will fall in the interval 83 to 87?
 - d) How large a sample is needed to be 95 percent sure that the sample mean will fall in the 83 to 87 interval?
 17. A large appliance manufacturer needs a current and accurate estimate of the retail sales of his appliances as an aid in production planning. Accordingly, the manufacturer plans to take a random sample of retail outlets and obtain sales on a monthly basis.
To aid in planning the survey, a preliminary sample of 60 retail outlets is selected. The results are $n = 60$, $\Sigma X = 1,104$, $\Sigma X^2 = 22,034$, where X is the appliance sales (units) by store in the past month.
-

- a) The manufacturer desires that the survey estimate of the mean sales per store be accurate within ± 1 appliance at the 95 percent level. How large must the total sample size be to achieve this precision?
- b) The cost of the survey is estimated at \$2,000 plus \$40 per store sampled. What is the cost of the survey designed in part a?
- c) Assume that the manufacturer distributes through 28,000 retail outlets. What will be the sampling error associated with the estimate of total monthly sales of appliances? (See hint given in question 12 above.)

SELECTED READINGS

Selected readings for this chapter are included in the list that appear on page 314.

12. TESTS OF HYPOTHESES

WE CAN MAKE a statistical inference either by estimating that the population mean (or other parameter) lies within a certain *confidence interval* or by *testing a hypothesis*. The sampling error $\sigma_{\bar{x}}$ is used in either case. Confidence intervals were considered in Chapter 11. In testing a hypothesis we first set up a hypothesis concerning the true population value of the mean μ , or some other parameter. Then we decide on the basis of a random sample whether to accept or reject this hypothesis. If the sample value is close to the hypothetical value, we accept the hypothesis; otherwise we reject it.

In the "classical" theory of statistical inference described in this chapter, one makes a decision either to accept or reject a hypothesis on the evidence of sample information alone. In Chapters 15 and 16 we will extend the analysis to include the judgment of the decision maker and the economic payoffs involved, using the "Bayesian" approach to arrive at an optimal decision.

The test of hypothesis approach is also useful in business and the social sciences for *reporting* purposes. In this sense, it serves to describe the sampling error associated with a given sample and to describe how likely the sample result could have occurred by chance alone.

An Example

Consider a specific example: In the manufacture of safety razor blades the width is obviously important. Some variation in dimension must be expected due to a large number of small causes affecting the production process. But even so, the average width should meet a certain specification. Suppose that the production process for a particular brand of razor blade has been geared to produce a mean width of 0.700 inches. Production has been underway for some time since the cutting and honing machines were last set, and the production manager wishes

to know whether the mean width turned out is still 0.700 inches, as intended.

This may be treated as a problem in statistical inference. It would be possible, of course, actually to measure all of the hundreds of thousands of blades turned out and to ascertain the mean width directly. But this would be expensive and very time-consuming. A better alternative would be to reason in terms of a sample. The statistical population of blade widths covers *all* the blades coming from the production line in the future under given technical controls. Since the production process was initially set up to give a mean width of 0.700 inches, the statistical hypothesis is posed that the true mean of this population is 0.700 inches. But the process could have gotten a little out of line, and management wishes to know whether 0.700 inches is still the mean width of all blades.

Accepting the Hypothesis. We have posed the hypothesis that the mean width of razor blades is 0.700 inches. This is written in symbols as $\mu_h = 0.700$, where μ_h is the hypothesized mean. The hypothesis seems reasonable since the machine was adjusted to this width. Suppose we draw a simple random sample of 100 blades from the production line. We measure each of these carefully and find the mean width of the sample to be 0.7005 inches. The standard deviation in the sample turns out to be 0.010 inches. That is,

$$\begin{aligned}n &= 100 \\ \bar{X} &= 0.7005 \text{ inches} \\ s &= 0.010 \text{ inches}\end{aligned}$$

For the hypothesis $\mu_h = 0.700$ to be true, the sample mean $\bar{X} = 0.7005$ inches would have to be drawn from the sampling distribution of all possible sample means whose overall mean is 0.700 inches.

Now, the important question arises: If the true mean of the population *really* were 0.700 inches, how likely is it that we would draw a random sample of 100 blades and find their mean width to be as far away as 0.7005 inches or farther? In other words, what is the probability that a value could differ by 0.0005 inches or more from the population mean *by chance alone*? If this is a high probability, we can accept the hypothesis that the true mean is 0.700 inches. If the probability is low, however, the truth of the hypothesis becomes questionable.

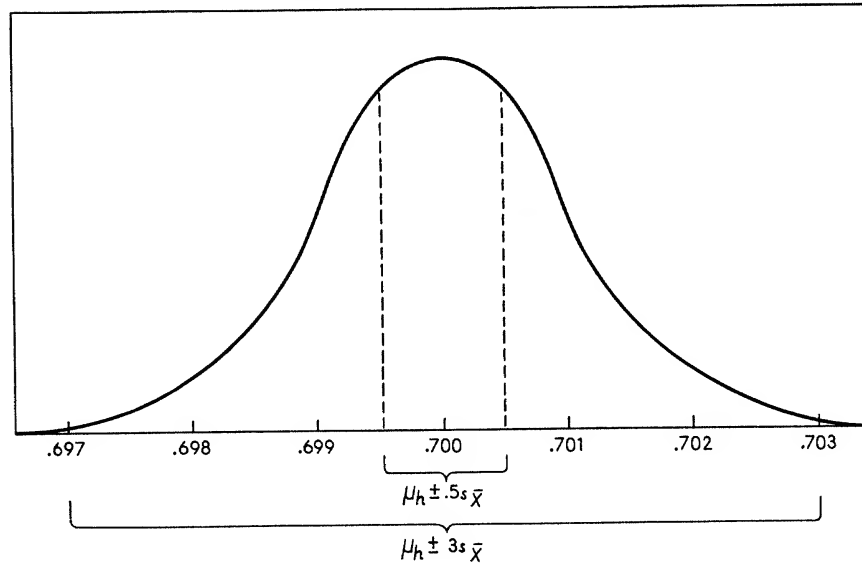
To get at this question, compute the standard error of the mean from the sample:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{.010}{\sqrt{100}} = 0.001 \text{ inches}$$

Since the difference between the hypothetical mean and the observed sample mean is 0.0005 inches, and the standard error of the mean is 0.001 inches, the difference is equal to 0.5 standard errors. By consulting Appendix D, we find that the area within this interval around the mean of a normal curve is $0.19 \times 2 = 38$ percent, so that $100 - 38 = 62$ percent of the total area falls *outside* this interval. (See dashed lines in Chart 12-1.) If 0.700 inches were the true mean, therefore, we

Chart 12-1

SAMPLING DISTRIBUTION OF MEANS OF
RAZOR-WIDTH SAMPLES OF SIZE 100
(Hypothetical Mean = 0.700 Inches)



should nevertheless expect to find that about 62 percent of all such possible sample means would, *by chance alone*, fall as far away as $0.5s_{\bar{x}}$ or farther. Therefore, the *probability* is 62 percent that our particular sample mean could fall this far away.

Remembering that we had substantial reason to accept the hypothesis in the first place—the process having been adjusted to yield a population mean of 0.700 inches—we should continue to hold to the hypothesis and attribute to mere chance the appearance of a 0.7005 inches mean in a single random sample of 100 blades.

Rejecting the Hypothesis. Later, after production has gone on for some time, the query again arises: Is it reasonable to believe that the true mean width of blades produced remains 0.700 inches? Since the

process was adjusted to yield that figure, the hypothesis still seems reasonable. We could then test it by taking another random sample of 100 blades. This time the standard deviation is still 0.010 inches, so the standard error of the mean is still 0.001 inches, but the mean is now 0.703 inches.

In order to test the hypothesis that the true mean of the population is 0.700 inches, we again go through the same line of reasoning. If the true population mean really were 0.700 inches, how likely is it that we should draw a random sample of 100 blades and find their sample mean to be as far away as 0.703 inches?

Since the difference between the hypothetical mean of 0.700 inches and the actual sample mean of 0.703 inches is 0.003 inches, and the standard error of the mean is 0.001 inches, the difference is equal to three standard errors of the mean (i.e., $0.003/0.001 = 3$).

Now, if 0.700 inches really were the population mean, we know from Appendix D that 99.7 percent of all possible sample means, for random samples of 100, would fall within three standard errors around 0.700 inches. (See wide bracket in Chart 12-1.) Hence, the probability is only 0.3 percent that we would get a sample mean falling as far away as ours does.

We have two choices:

1. We may continue to accept the hypothesis (i.e., leave the production process alone), and attribute the deviation of the sample mean to chance.
2. We may reject the hypothesis as being inconsistent with the evidence found in the sample (hence, correct the production process).

Either of two things is true: (1) the hypothesis is correct, and an exceedingly unlikely event has occurred by chance alone (one which would be expected to happen only 3 out of 1,000 times); or (2) the hypothesis is wrong. We have to make a decision between the two.

In this case, if we decide on the sample information alone, we would probably make choice (2) and conclude that the mean width of blades from that production line was not really 0.700 inches. We would reject the hypothesis as being inconsistent with the evidence found in the sample. We would then be wrong only when the hypothesis was actually true and by chance alone a sample mean fell as far away as three standard errors. But on the average this would occur only 3 in 1,000 times.

The Choice between Accepting and Rejecting the Hypothesis.

Ultimately, in our example, the choice between letting the production process alone and the alternative of stopping the process to make adjustments depends upon other factors in addition to the sample evidence. The cost of incorrectly stopping the process and the cost of allowing a faulty process to continue are certainly relevant. In addition, the past history of this manufacturing process also influences the choice. If the process rarely goes out of adjustment, we would be more inclined to attribute a far-out sample mean to chance than we would if the process frequently went out of adjustment. The problems of incorporating prior judgment and economic losses are discussed in Chapter 15. In addition, Chapter 25 on quality control discusses in detail the control of production processes.

The hypothesis testing analysis, however, is itself helpful. It deals with the evaluation of the sample and the conclusions that may be drawn from that evidence *alone*. In a sense, it is a method of reporting on the sampling error for a given sample. Rejecting the hypothesis means that the sample evidence is strongly against the hypothesis. Accepting the hypothesis means that the evidence is not in disagreement with the hypothesis.

A legal analogy may help in understanding the reasoning involved. In a sense, the hypothesis is on trial and is considered innocent until proved guilty. The evidence is found in the random sample. Before the hypothesis is condemned, the evidence must prove it guilty—not with absolute certainty, but beyond reasonable doubt. The particular form which the evidence takes is the probability that a value as different as the sample mean could have been drawn if the hypothesis were true. If this probability is high, we can accept the hypothesis. On the other hand, if this probability is low, the hypothesis is doubtful. The lower the probability, the progressively greater is the doubt that the hypothesis could be correct. Finally, if the probability is so low that it appears unacceptable to believe that a value as different as the sample mean could have arisen solely by chance, the hypothesis is rejected. It is judged guilty beyond reasonable doubt.

In the first example just considered, the probability was quite high (62 percent) that a discrepancy of 0.0005 inches could be attributed to mere chance. Therefore, we accepted the hypothesis, particularly since we had pretty good reason to believe in it before the sample was drawn. We could easily view the hypothetical mean of 0.700 inches as compatible with the findings of the sample *and* the operations of chance. But in the second example given ($\bar{X} = 0.703$ inches), the probability was so

low (0.3 percent) that such a large difference could arise by chance, that the hypothesis ($\mu_h = 0.700$ inches) was rejected as being untrue.

It is important to note that while rejection of a hypothesis implies that the hypothesis is false, *acceptance of a hypothesis does not necessarily prove that the hypothesis is true*. It may be that the hypothesis is in fact false (i.e., that the true mean μ differs from μ_h) but the sample does not have sufficient precision (i.e., the sampling error is too large) to be able to detect the difference. We shall examine this possibility in more detail shortly.

TYPE I AND TYPE II ERRORS

Understandably, the question can be raised: What critical value should we select for the probability of getting the observed difference ($\bar{X} - \mu_h$) by chance, above which we should accept the hypothesis and below which we should reject it? This value is called the *critical probability* or *level of significance*. The answer to this question is not simple, but to explore it will throw further light on the nature and logic of statistical inference.

Only four possible things can happen when we test a hypothesis. We may be wrong because we:

1. Reject a true hypothesis (a Type I error), or
2. Accept a false hypothesis (a Type II error).

Or, we may be right because we:

3. Accept a true hypothesis, or
4. Reject a false hypothesis.

The types of errors noted as possibilities 1 and 2, respectively, are known either as Type I and Type II errors or as errors of the first kind and errors of the second kind.

Type I Errors

In a long run of cases in which the hypothesis is in fact *true* (although we do not know it is true, for otherwise there would be no need to test it), we will necessarily either be wrong as in 1 or right as in 3. That is to say, if we make an error it will have to be Type I. Suppose we should adopt 5 percent as the critical probability, accepting the hypothesis when the probability of getting the observed difference by chance exceeds 5 percent and rejecting the hypothesis when this probability

proves to be less than 5 percent. This amounts to the decision to accept the hypothesis when the discrepancy of the sample mean is less than 1.96 standard errors, and to reject the hypothesis when the discrepancy is more than 1.96 standard errors. Using this value as the critical probability, we would expect to make a Type I error 5 percent of the time. This is because even when the hypothesis is true, 5 percent of all possible sample means still lie farther away than 1.96 standard errors. And whenever by chance we get one of these, and the hypothesis is true, we would make the mistake of rejecting a true hypothesis.

Or, we might choose 1 percent as the critical probability, which would correspond to a discrepancy between hypothesis and the sample mean equal to 2.58 standard errors. When the hypothesis is in fact true, only 1 percent of all possible sample means would lie farther away than 2.58 standard errors. We would make a Type I error only when by chance alone we happened to draw one of these. Which is to say, we would now make an error of the first kind only 1 percent of the time.

Clearly, then, the proportion of cases in which we would make an error of the first kind, that of rejecting a true hypothesis, can be made as small as we wish simply by reducing the value for the critical probability. In fact, *the percentage of cases in which we would expect to make an error of the first kind is precisely equal to the critical probability adopted.*

Just Significant Probability Level. In many studies, the critical probability is used to describe the statistical significance of a sample result. For example, an economist collects some data on, say, interest rates and the demand for money. He hypothesizes some relationship and wishes to see if the data support his thesis. He tests the hypothesis to rule out the alternative that the observed relationship occurred by pure chance. He then reports his sample result as "significant at the 1 percent level." Such a statement is a report to the reader that has the following meaning: (1) if we were to set up a statistical hypothesis (and the particular hypothesis is either stated or is obvious from the context of the problem); and (2) if we were to test this hypothesis using a critical probability (or significance level) of 1 percent; then (3) we would reject the hypothesis and rule out a chance relationship.

Significance levels (critical probabilities) of 10, 5, 1, and 0.1 percent are often used in reporting sample data. The smallest of these probability values is chosen at which the hypothesis can be rejected. In other words, the *just significant probability level* is reported.

To make this clear, suppose that the analyst in the razor blade example were reporting the results of a sample of 100 razor blades to a superior.

With a sample mean $\bar{X} = 0.703$ and a standard error $s_{\bar{X}} = 0.001$, the sample mean is 3 standard errors away from the hypothesized mean. The analyst might therefore describe the sample mean as "significantly different from 0.700 inches at the 1 percent level of probability." The use of a 1 percent critical probability would reject any sample mean outside $\mu \pm 2.58 s_{\bar{X}}$. Note that the sample result could *not* be described as significant at the 0.1 percent level, which would require a deviation of 3.28 standard errors. This use of the hypothesis testing procedure, therefore, is a reporting or communication technique. It is used in the same manner as a confidence interval to describe the sampling error associated with a given sample.

Type II Errors

So far we have concerned ourselves only with the first kind of error. But there is also the second kind—the possible error of accepting a false hypothesis. The lower the value we set for the critical probability, in general the fewer the hypotheses we will reject. But the chances are then increased of accepting more hypotheses which are false. We can buy safety in one direction only at the expense of danger in the other.

Unfortunately, it is impossible to predict in general the percentage of times we should expect to commit an error of the second kind on the basis of any particular value adopted for the critical probability. The reason for this is that the chance of accepting a false hypothesis depends also upon *how* false the particular hypothesis happens to be. Remember that sample means tend to cluster around the true means of the populations from which they are drawn. If the hypothetical mean is far away from the true mean, it is unlikely that a sample mean will be drawn which appears consistent with the hypothesis. If the hypothetical mean is false but not far from the mark, an error of the second kind is much more likely to be made.

In a long run of instances in which hypotheses are actually false some will be farther from the true mean than others. Therefore, it is impossible to predict in general the probability of accepting false hypotheses. We can appreciate, however, that the chances of accepting false hypotheses are increased as fewer hypotheses are rejected due to the use of a lower value for the critical probability. The problem of balancing Type I against Type II errors is discussed below.

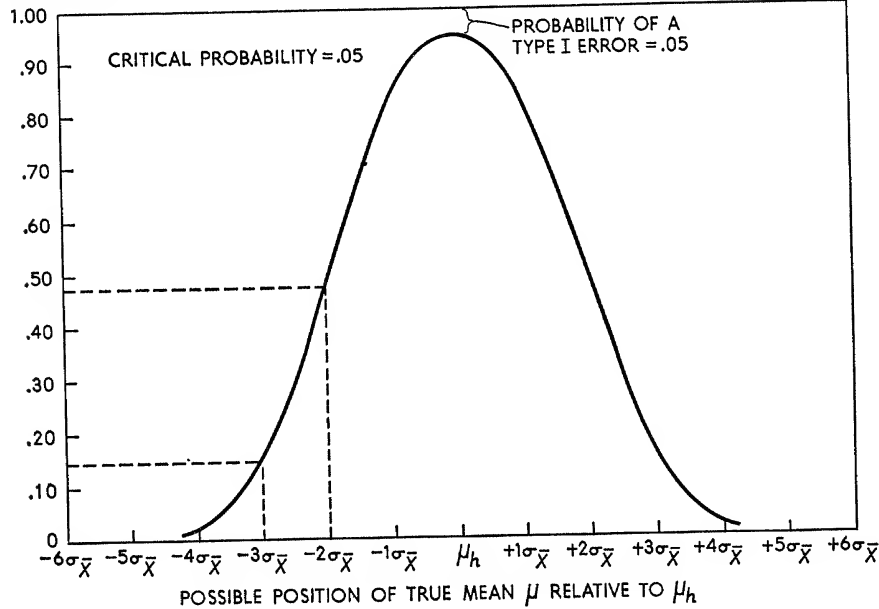
Operating Characteristic Curves. The exact probability of making a Type II error depends upon how far the true mean μ of the population is away from μ_h , the hypothetical mean. This can best be illustrated by an *operating characteristic curve* or *OC curve*, as shown in Chart 12-2.

Chart 12-2

PROBABILITY OF ACCEPTING THE HYPOTHESIS
FOR ALL POSSIBLE ALTERNATIVE MEANS
(Operating Characteristic Curves)

A

PROBABILITY OF
A TYPE II ERROR;
ACCEPTING THE
HYPOTHESIS

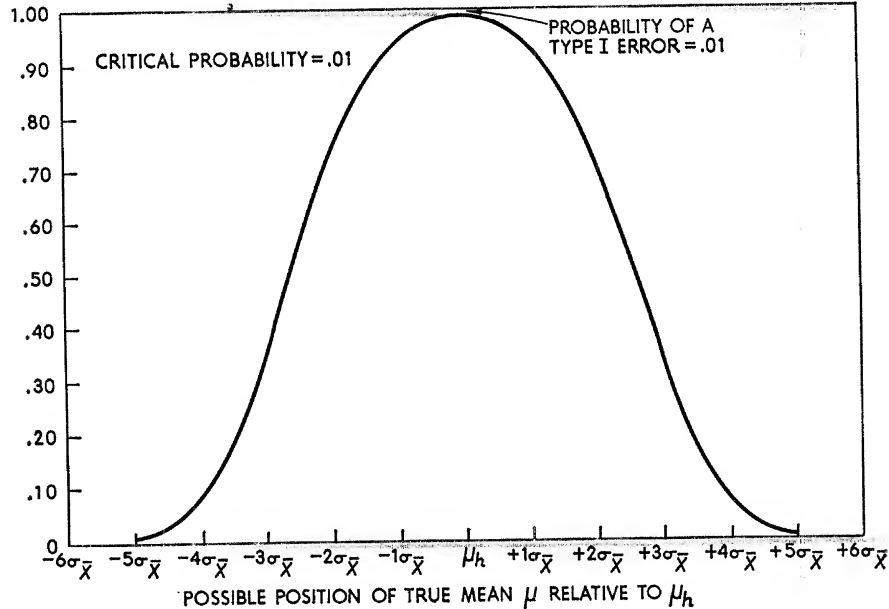


The vertical scale of Chart 12-2 shows the probability of committing a Type II error (i.e., accepting the hypothesis when it is false). The horizontal scale shows all possible values for the true mean of the population, relative to the hypothetical mean μ_h . Thus, if the true mean were one standard error less than μ_h , it would be at the point $-1\sigma_{\bar{x}}$ on the horizontal axis. Panel A represents the use of a critical probability of 0.05, and panel B a critical probability of 0.01. In either case, the probability of a Type II error can be found for any possible value of the true mean. Thus, in Chart 12-2A, if the true mean were three standard errors below the hypothetical mean ($-3\sigma_{\bar{x}}$), the probability of a Type II error would be 0.15, as shown by a dashed line. Similarly, if the true mean were two standard errors below the hypothetical mean ($-2\sigma_{\bar{x}}$), the probability of a Type II error would be 0.48.

When the true mean is exactly at the hypothetical mean ($\mu = \mu_h$), a Type II error is impossible. Then the distance from the top of the curve

Chart 12-2 (Continued)
B

PROBABILITY OF
A TYPE II ERROR:
ACCEPTING THE
HYPOTHESIS



to 1.0 represents the probability of a Type I error. Thus, since 0.95 is the probability of accepting the hypothesis when $\mu = \mu_h$, 0.05 is the probability of rejecting it (when it is true), that is, of committing a Type I error.

Balancing Type I against Type II Errors

In testing hypotheses, we face two dangers: that of rejecting a true hypothesis and that of accepting a false hypothesis. The danger of committing a Type I error can be made as low as we please by reducing the value chosen for the critical probability; but this can be done only at the expense of increasing the danger of committing a Type II error. This can be seen by comparing the two curves in Chart 12-2. The probabilities on Chart 12-2B (with the more stringent critical probability of 0.01) are higher at every point than on Chart 12-2A.

The "classical" approach to statistical inference would leave the

balancing of these risks and the determination of the critical probability to the judgment of the analyst. In the razor-blade example a Type I error would mean falsely condemning the accuracy of a production process which was in fact operating as intended. A Type II error would mean continued production of a product which in fact was not meeting specifications. The economic penalty of the Type I error might be an expensive shutdown to look for a nonexistent trouble. The economic consequences of the Type II error might be the loss of consumer goodwill as the customers later found the product unsatisfactory. (They might get razor burn with undue frequency, or find that the average blade did not fit into the razor.) With these potential economic consequences in mind, it would be up to management to set the value for the critical probability where, in its judgment, the best compromise is reached between risks of incurring the two types of errors.

In the "Bayesian" approach to statistical inference, the economic risks as well as the judgment of the decision-maker are included in a formal decision-making procedure. This approach is the subject of Chapters 15 and 16.

Effect of Sample Size on Probability of Errors

So far the discussion of hypothesis testing has been in terms of some particular size of sample. So long as a given sample size is assumed, the risk of a Type I error can only be reduced at the expense of increasing the risk of a Type II error. There is, however, a way of reducing the chance of accepting a false hypothesis without at the same time increasing the chance of rejecting a true hypothesis. By taking a larger sample the *combined* chance of committing either error can be reduced.

As the size of the sample drawn is increased, \bar{X} will tend to fall closer to the actual value for μ , since $s_{\bar{X}}$ is decreased. With any particular value for the critical probability, Type I errors will be made with the same relative frequency, whatever the sample size. But as \bar{X} is pulled in closer to μ (as is the tendency in taking a larger sample), \bar{X} will in fewer instances appear consistent with a value other than μ , that is, with a false hypothesis regarding μ .

Thus, by taking a larger sample, the chance of a Type II error (accepting a false hypothesis) is reduced, while the chance of rejecting a true hypothesis can be held constant by using the same value for the critical probability. The combined chance of error will be smaller if we can reduce one component while we hold the other chance component constant. Just as we might expect, fewer overall mistakes of statistical inference will be made the larger the size of sample used.

TWO-TAILED TESTS VERSUS ONE-TAILED TESTS

In the form of testing hypotheses so far discussed the probability has been calculated of getting a discrepancy as large as or larger than that observed by adding together the two "tails" of the sampling distribution beyond the number of standard errors corresponding to $(\bar{X} - \mu_h)$. This is referred to as "testing in both directions" or as a "two-tailed test."

Two-Tailed Tests

In the first example the probability of 62 percent was attached to the likelihood of getting a discrepancy as large as or larger than that observed ($0.5s_{\bar{X}}$), regardless of the sign of the discrepancy, that is, whether it might have arisen by $\bar{X} \geq 0.7005$ inches or $\bar{X} \leq 0.6995$ inches. In the second example, the probability of 0.3 percent was calculated for the chance of getting a difference equal to or exceeding that observed ($3s_{\bar{X}}$), whether that difference be above or below 0.700 inches.

There are three related reasons for testing in both directions when testing a single numerical value (such as 0.700 inches) as being the true mean of the population:

1. The hypothesis is in theory formed before the sample is drawn; hence, we don't know in advance whether the observed discrepancy between μ_h and \bar{X} will have a positive or a negative sign.
2. An observed discrepancy of any particular size would be equally harmful to the hypothesis, whether it had a positive or a negative sign.
3. A hypothesis must not be rephrased to incorporate any of the information found in the very same sample which is used to test it.

The last point requires a bit of expansion. The hypothesis that the mean width of blade is 0.700 inches is a *single-valued hypothesis*; it says not greater than that, not less than that. If, on finding \bar{X} equal to 0.703 inches, we had calculated only the probability of getting by chance a sample mean as large as or larger than 0.703 inches, we would have subtly shifted our initial hypothesis to the hypothesis that the population mean is *not greater than* 0.700 inches. Implicitly, we would have wound up testing a different hypothesis than the one intended, and simply because of the sign of the discrepancy which was found after the sample had been drawn.

In the razor-blade case it seemed quite appropriate to test the

single-valued hypothesis of 0.700 inches, that is, to test in both directions, since presumably we would be just as concerned about blades being too wide as being too narrow.

One-Tailed Tests

In other cases, however, it might be appropriate to test in one direction only; that is, to test what can be called a *multivalued hypothesis*.

If we were concerned with the strength of parachute cords, we would not be worried about their being too strong; we would worry only about their being too weak. If for safety's sake they were designed, let us say, to have a mean breaking point of 1,000 pounds, we would be interested in the hypothesis that the true population mean was *not less than* 1,000 pounds. Correspondingly, we would test the multivalued hypothesis that the true mean had a value of 1,000 pounds or some larger value.

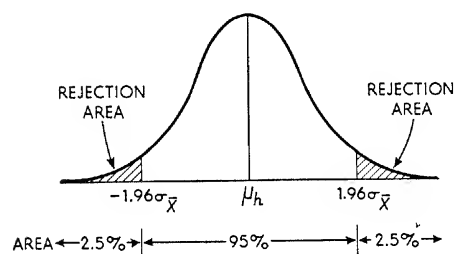
Should a sample mean greater than 1,000 pounds be found in a random sample drawn, it would immediately be accepted as consistent with the hypothesis. Only if \bar{X} should be less than 1,000 pounds would a question arise concerning the validity of the hypothesis. It would then be appropriate to ask the question: If the mean of the population truly were 1,000 pounds or more, what is the probability of getting by chance a sample mean which falls below 1,000 pounds by as much as the one observed? That is to say, the particular sign of the observed difference now would have a bearing on the truth or falsity of the hypothesis as stated. It is appropriate in this case to test in but one direction, that is, in terms of the probability of getting by chance a sample mean which lies below 1,000 pounds by an amount equal to or greater than that observed.

One important change is made when applying a one-tailed test instead of a two-tailed test, namely, the multiple of the standard error which corresponds to any given critical probability. In a two-tailed test,

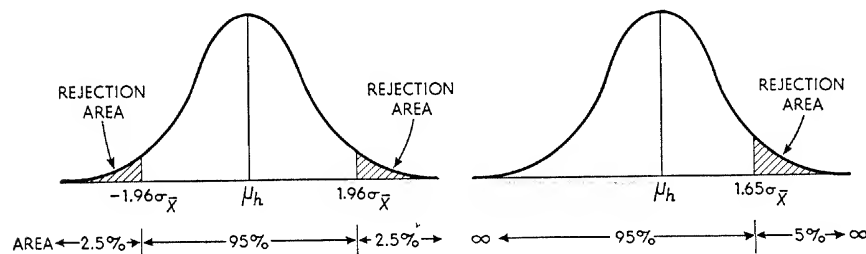
Chart 12-3

AREAS OF REJECTION—5 PERCENT CRITICAL PROBABILITY

A. TWO-TAILED TEST



B. ONE-TAILED TEST



$1.96\sigma_{\bar{x}}$ corresponds to a 5 percent critical probability, whereas $1.65\sigma_{\bar{x}}$ is the multiple of the standard error associated with 5 percent in a one-tailed test. When testing in both directions, $2.58\sigma_{\bar{x}}$ goes with 1 percent as the critical probability. But for testing in a single direction, the similar combination is $2.33\sigma_{\bar{x}}$ and 1 percent. These can be read from Appendix D for various areas under the normal curve.

For a 5 percent critical probability under a two-tailed test and one-tailed test, respectively, see Chart 12-3.

TESTS OF DIFFERENCES BETWEEN ARITHMETIC MEANS

We now consider another important aspect of statistical inference, namely, tests of the significance of differences between sample means. This phase is concerned with the following problem: Given an observed difference between the means of two random samples, each drawn from a different population, is this difference to be taken as signifying a real difference between the true means of the populations involved?

To handle this problem it is necessary to introduce the concept of a new sampling distribution, the sampling distribution of *differences* between means. We can think of this distribution as being formed in the following manner.

On the basis of random sampling from two separate populations, the sampling distributions of the arithmetic means \bar{X}_1 and \bar{X}_2 would be formed. Each of these sampling distributions is of the same type we have been discussing.

Now imagine that from each of these sampling distributions a sample mean is drawn at random and that the difference between this pair of sample means is noted. Then a second pair of sample means is selected at random, each from its own sampling distribution. The difference between this second pair almost certainly would be different from that found between the first pair, due to chance alone. We can imagine the process carried on indefinitely. Then we would have an indefinitely large number of values representing the differences between all possible pairs of sample means which could be drawn at random from their respective populations. These differences would form a theoretical distribution known as the sampling distribution of the difference between two means.

We know the following things about this new distribution.

1. The sampling distribution of differences tends to be normal; which is to say that differences between pairs of sample means
-

- will be normally distributed, provided that the sample size is large.
2. The mean of the distribution of differences will be the true difference between the population means ($\mu_1 - \mu_2$). This follows from the proposition that the mean of the differences between any two series of values is equal to the difference between their respective means.
 3. The standard deviation of the distribution of differences may be estimated by the formula

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}$$

In this formula $s_{\bar{x}_1}$ is the standard error of the mean for the sampling distribution of \bar{X}_1 and $s_{\bar{x}_2}$ is the similar measure for the sampling distribution of \bar{X}_2 . The value $s_{\bar{x}_1 - \bar{x}_2}$ is known as the *standard error of the difference between two means*.¹

With this important new sampling distribution in mind, we can carry forward our discussion of the present phase of statistical inference in terms of specific examples.

Suppose a trucking firm is testing two brands of truck tires for their wearing ability in order to decide if one brand has greater average mileage than the other. One hundred tires of brand No. 1 are put on the firm's trucks and the mileage records are kept until the tires are worn out; similarly, 144 tires of brand No. 2 are put on trucks and the mileage is recorded. Both brands of tires are placed at random on the firm's trucks to guard against any systematic bias because of characteristics or usage of certain trucks.² (A difference in sample size is used in this example merely to emphasize that the two samples need not be equal in size for this method to be applicable.) The following means and standard deviations result (the subscripts referring to the brand number):

<i>Tire Brand No. 1</i>	<i>Tire Brand No. 2</i>
$n_1 = 100$	$n_2 = 144$
$\bar{X}_1 = 37.4$ thousands of miles	$\bar{X}_2 = 36.8$ thousands of miles
$s_1 = 5.1$ thousands of miles	$s_2 = 4.8$ thousands of miles

¹ In this discussion s represents the standard error estimated from a sample; if the true population value were known, the symbol σ would be used, with appropriate subscript.

The variance (s^2) of the difference is the sum of the variances of the individual means. As a graphic check, the standard error of each mean can be laid off as a side of a right triangle; then the standard error of the difference can be read off as the hypotenuse.

² A better statistical design, perhaps, would call for putting both brands on the same truck to reduce differences due to truck characteristics and usage. For more on this technique of pairing observations, see pages 124–127 of the Dixon and Massey reference listed on page 314 and problem 14 at the end of this chapter.

The test gives tire brand No. 1 an advantage of $\bar{X}_1 - \bar{X}_2 = 0.6$ thousand miles in average mileage. Nevertheless, because we are quite aware of chance variations that may occur in random sampling, we do not immediately jump to the conclusion that brand No. 1 is longer wearing than brand No. 2. We are led to wonder if the difference in mean mileage observed in the samples arose by chance or whether there is in fact a difference in average mileage between all tires of brand No. 1 and all tires of brand No. 2. That is to say, we wish to know if the observed difference between the sample means indicates a real difference between the means of the two populations.

The Null Hypothesis

Our manner of solving this problem is to set up and test the so-called "null hypothesis." This means that we pose the hypothesis that there is *no* difference in average mileage between brand No. 1 and brand No. 2, and then proceed to test that hypothesis against the evidence found in the samples.

The null hypothesis states that the mean of the sampling distribution of differences is equal to zero. This is because the mean of the sampling distribution of differences is known to be $(\mu_1 - \mu_2)$, and the hypothesis is that there is no difference between these population means.

The observed difference of 0.6 thousand miles between the two random sample means is, in effect, one observation drawn at random from the sampling distribution of all possible differences between pairs of random sample means. We can therefore ask the question: If the mean of the sampling distribution of differences really were zero, what is the probability that we would get a difference between two sample means at least as large as 0.6?

Since the sampling distribution from which 0.6 came tends to be normal, we can answer this question as soon as we know the value for the standard error of the difference between means. This is computed as follows:

$$\begin{aligned}s_{\bar{x}_1} &= \frac{5.1}{\sqrt{100}} = 0.51 & s_{\bar{x}_2} &= \frac{4.8}{\sqrt{144}} = 0.40 \\s_{\bar{x}_1 - \bar{x}_2} &= \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2} \\&= \sqrt{(0.51)^2 + (0.40)^2} \\&= \sqrt{.4201} \\s_{\bar{x}_1 - \bar{x}_2} &= 0.65\end{aligned}$$

Accepting the Null Hypothesis. Thus, it turns out that the observed difference between the sample means is less than one standard error of the difference ($0.6/0.65 = 0.92$ standard errors, to be exact). If the true difference between the population means really was zero, the probability is nevertheless 36 percent that a difference at least as large as 0.6 thousand miles would appear by chance. It would appear that there is no compelling evidence to be found in the samples that a real difference exists in average mileage between the two brands. In this case it is said that the difference between the sample means is too small to be significant—that is, too small to signify an indisputable difference between the population means.

Rejecting the Null Hypothesis. Let us take the same case again, but assume \bar{X}_1 had come out 38.6 instead of 37.4 thousand miles. Now the observed difference between the sample means is $38.6 - 36.8 = 1.8$ thousand miles. This in turn is equal to 2.8 standard errors of such differences (i.e., $1.8/0.65 = 2.8$). Since 2.8 is greater than the 2.58 standard errors associated with a 0.01 probability level, the observed sample difference is significant at the 0.01 level.

Actually, if there really were no difference between μ_1 and μ_2 , the probability of getting an observed difference equal to or greater than 2.8 standard errors in either direction would be only 0.5 percent. It appears highly unlikely, therefore, that the difference between the means of the samples could have appeared solely by chance in this case. The null hypothesis may very well be rejected.

The Choice between Acceptance and Rejection. In the first instance above, a difference in the sample means of 0.6 thousand miles or more could occur by chance 36 percent of the time. Most observers, on the basis of the sample alone, would accept the hypothesis. Such an acceptance would imply either (1) that there was no difference in mean wearing ability of the two brands of tires and the observed sample difference was due to chance or (2) that there was a difference but the samples were too small to detect the difference. On the other hand, a difference in sample means of 1.8 thousand miles is significant at the 0.01 level and strongly indicates a real difference in mean wearing ability.

What would be the conclusion if, for example, the difference in the sample means were 1.0 thousand miles or 1.5 standard errors ($1.0/0.65 = 1.5$)? The probability of a difference in sample means this large or more is 13 percent. In such a case, we conclude that the sample gives some evidence that one tire is longer wearing than the other on the average, but the possibility that the sample result is due to

chance cannot be ruled out. In other words, on the basis of the sample alone, the results are inconclusive.

If some action must be taken—e.g., which tire to purchase—evidence other than the sample would be included in the decision analysis. The past reputation of the tire manufacturers, the prices of the two brands, as well as the savings associated with longer tire wear should be considered. In the “classical” statistical approach, these factors should be incorporated in the determination of the appropriate Type I and II error probabilities. In the “Bayesian” approach, these factors are explicitly included in the decision-making procedure (see Chapters 15 and 16).

Confidence Intervals for the Difference between Sample Means

Rather than testing the hypothesis that there is no difference in population means, we may wish to estimate the actual difference between the means. The procedure, in principle, is identical with that employed earlier in estimating the mean of a population on the basis of the mean of a random sample drawn from that population. The only difference is that the sampling distribution of differences (and its associated measures) is employed in forming the appropriate confidence intervals in the present case.

We wish to estimate $(\mu_1 - \mu_2)$, which is known to be the mean of the sampling distribution of differences. From this sampling distribution we have one observation $(\bar{X}_1 - \bar{X}_2)$, based upon random sampling. Then 68 percent of such observations would be expected to lie within $s_{\bar{X}_1 - \bar{X}_2}$ of the mean difference; 95 percent would be expected to lie within $1.96s_{\bar{X}_1 - \bar{X}_2}$ of $(\mu_1 - \mu_2)$ etc. Consequently, we should have a 68 percent degree of confidence that an interval specified as $(\bar{X}_1 - \bar{X}_2) \pm s_{\bar{X}_1 - \bar{X}_2}$ would include the value $(\mu_1 - \mu_2)$ and a 95 percent degree of confidence that the interval $(\bar{X}_1 - \bar{X}_2) \pm 1.96s_{\bar{X}_1 - \bar{X}_2}$ would include the true difference between the population means.

In the second example above, the observed difference is 1.8 thousand miles; with a standard error of 0.65 thousand miles. We may estimate, therefore, that the true difference between the population means lies within the interval 1.8 thousand miles \pm 1.3 thousand miles (i.e., 1.96 times the standard error) and hold a 95 percent degree of confidence that our estimate is correct. The 95 percent confidence limits are then 0.5 thousand miles and 3.1 thousand miles for the superiority of tire No. 1 over tire No. 2 as regards average mileage.

If the confidence interval based upon $\pm 3s_{\bar{X}_1 - \bar{X}_2}$ is computed to give a degree of confidence of 99.7 percent that the true difference is located within its boundaries, the confidence limits work out to be minus 0.15

thousand miles to 3.75 thousand miles for the difference between brands No. 1 and 2 in average mileage. This result—the appearance of the negative sign for the lower limit of the confidence interval—might puzzle the student, but it really need not. All it means is that for us to be 99.7 percent confident that we have located the real difference in average mileage between the two brands we should have to grant that superiority *might* lie to a small extent with brand No. 2.

SUMMARY

We can make a statistical inference either by constructing a *confidence interval* (as described in Chapter 11) or by *testing a hypothesis*. In the latter case we set up a hypothesis regarding the value of the parameter—say, the mean. If the sample mean is close to the hypothetical mean, we accept the hypothesis; otherwise we reject it.

In the case of the razor-blade machine that was set to produce blades of average width 0.700 inches, a sample of 100 blades was tested, with $\bar{X} = 0.7005$ inches and $s = 0.010$ inches, so $s_{\bar{X}} = s/\sqrt{n} = 0.001$ inches. Since the sample mean was only 0.5 standard errors away from the hypothetical mean, the probability was 62 percent of getting such a discrepancy by chance, so the hypothesis was accepted. In a second trial, however, with $\bar{X} = 0.703$ inches, the hypothesis ($\mu_h = 0.700$ inches) was rejected since it was quite unlikely that such a discrepancy could occur by chance alone. A reasonable hypothesis is usually accepted unless the probability is quite low (say, under 5 percent or even 1 percent) that the discrepancy of the sample value could be attributed to chance. The problem is where to set this *critical probability* below which we will reject the hypothesis. Rejection of a hypothesis indicates a belief that the hypothesis is false. Acceptance of a hypothesis, however, does not necessarily prove that the hypothesis is true. It may be that the sample is too small to detect a significant difference.

We can make two types of errors in testing hypotheses:

1. Type I: rejecting a true hypothesis.
2. Type II: accepting a false hypothesis.

We can easily control the chance of making a Type I error, since this equals the critical probability that is set in advance. Unfortunately, for a given size of sample, we can reduce the chance of making a Type I error only at the cost of increasing the risk of making a Type II error. The chance of making the latter error is unknown, since it depends on how far the hypothetical mean is away from the true mean.

By taking a larger sample, the combined chance of making either error can be reduced. In particular, if the critical probability is held constant, the chance of a Type I error also remains constant, in a larger sample, but the chance of a Type II error is reduced.

An operating characteristic or *OC* curve shows the probability of making a Type II error (that is, accepting the hypothesis when it is false) for a given critical probability, depending on how far the true mean is from the hypothetical mean. The farther these means are apart, the smaller is the probability of a Type II error.

The critical probability used in hypothesis testing is determined, in the "classical" approach to statistical inference, by balancing the Type I and Type II errors. If a Type I error would be serious relative to a Type II error, the critical probability should be set relatively low. When the relative costs cannot be determined, critical probabilities are often set at the arbitrary values of 5 or 1 percent.

In the "Bayesian" approach to statistical inference (Chapters 15 and 16) the economic consequences as well as the prior judgment of the decision-maker are included with the sample in making a decision.

Business and economic studies often report a sample result as, for example, "significant at the 1 percent level." Such a statement describes the sampling error associated with a sample and indicates that an implied hypothesis would be rejected if a 1 percent critical probability were used. Significance levels of 10, 5, 1, and 0.1 percent are commonly used, and the smallest probability at which the hypothesis will be rejected is reported.

In testing hypotheses, we may make either a two-tailed or a one-tailed test. The *two-tailed test* takes into account the areas under both tails of the normal curve (Chart 12-3). It is appropriate in most practical situations because we are concerned with discrepancies *either* above or below the hypothetical mean. In case we are concerned only with discrepancies in one direction from the hypothetical mean, however, it is appropriate to use the *one-tailed test*, which takes into account only the area under one tail of the normal curve. The decision rule is then to reject the hypothesis if $(\bar{X} - \mu_h)/s_{\bar{X}}$ exceeds the following values:

<i>Critical Probability Chosen</i>	<i>Two-Tailed Test</i>	<i>One-Tailed Test</i>
5 percent	1.96	1.65
1 percent	2.58	2.33

We can also test whether the *difference* between two sample means signifies a real difference between the population means or whether the observed difference is merely due to chance. To do this, we find the

standard error of the difference (theoretically, the standard deviation of a distribution of differences between many pairs of sample means). This is computed from the standard errors of the individual means. Then we can test the *null hypothesis* (that there is *no* difference between the population means) by expressing the difference between the sample means as a ratio of their standard error. If this ratio is small, we accept the null hypothesis; otherwise we reject it, depending on the probability that the difference could be due to chance (from Appendix D), and balancing the consequences of Type I and II errors as before. We can also set up a *confidence interval* around the difference between the sample means, based on its standard error, as was done earlier.

PROBLEMS

1. Distinguish between:
 - a) Confidence intervals and tests of hypotheses.
 - b) Type I and Type II errors.
 - c) How to find the probability of Type I and Type II errors from an operating characteristic curve.
 - d) One-tailed and two-tailed tests.
 - e) Use of hypothesis testing for decision-making and for reporting.
 2. A grocery chain store adopts a policy of issuing trading stamps on all purchases. Prior sales had averaged \$15.50 per customer over the past year, with a standard deviation of \$4.80. At the end of a trial period with the new stamps, a random check of 400 customers shows average sales of \$16.30. Have the stamps increased the average sales?
 3. A machine, when in adjustment, produces parts that have a mean diameter of 0.300 inches with a standard deviation of 0.012 inches. A random sample of 36 parts yields a mean diameter of 0.297 inches. Is the machine probably still in adjustment or not? Give reasons.
 4. If we change the critical probability from 5 to 0.1 percent, what is the effect on:
 - a) The probability of rejecting a true hypothesis?
 - b) The probability of accepting a false hypothesis?
 5.
 - a) Suppose the null hypothesis is $\mu_h = 14.0$, $n = 25$, $\sigma = 2.0$, and the critical probability is 0.05. Using Chart 12-2, what would be the probability of a Type II error if the actual μ of the population were 15.0? If the actual μ were 14.5?
 - b) What would be the probability of a Type II error if the sample size were increased to 36 and the actual μ were 15.0? If the actual μ were 14.5?
 - c) What would be the probability of a Type II error for $n = 25$, if a 0.01 critical probability were to be used and the actual μ were 15.0? If the actual μ were 14.5?
-

6. The standard time for a certain assembly operation is 2.4 minutes. Jones has been observed and timed in this operation 32 times over the past two weeks with the following results: X = observed time in minutes for Jones to complete the assembly operation; $n = 32$, number of observations of Jones; $\bar{X} = 2.8$ minutes; $\Sigma X = 89.6$; $\Sigma X^2 = 320.63$.
- If the evidence is sufficiently strong that Jones is not meeting the standard on the average, then he is to be retrained. What conclusion can you draw from the sample result? What action should be taken?
7. A certain pneumatic tool is designed so that it should operate on a pressure of no more than 20 pounds per square inch. Management was receiving complaints from purchasers that the pressure necessary to operate the tools was in excess of the 20 pounds psi standard. To check this, 40 tools were selected from current production, and the operating pressure was checked on each under controlled conditions. The results were X = pressure in pounds per square inch to operate a given tool; $n = 40$; $\Sigma X = 740$; $\Sigma X^2 = 14,041$.
- Is a one- or two-tailed test appropriate in this situation?
 - What can you conclude from the statistical test of hypothesis?
 - Does your answer to *b* reply to the objection raised by the customers? Why or why not?
8. A manufacturer of incandescent lamps is testing to see if the average life of the lamps he is manufacturing is above or below the standard of 2,000 hours. To check, the manufacturer proposes to take a sample of 200 lamps and to determine the life of each. And he plans on using a 1 percent critical probability (two-tailed). From past experience, the standard deviation of the burning life of this type of lamp is known to be about 1,000 hours.
- What is the hypothesis?
 - What is the meaning of a Type I error in this situation? What is the probability of a Type I error?
 - Suppose that the true mean life deviates by 100 hours from the standard. What is the probability that the sample will be able to detect the difference?
 - Suppose that the true mean life deviates by 200 hours from the standard. What is the probability that the sample will be able to detect this difference?
 - Suppose that the true mean life differs from the standard by 150 hours. How large a sample would be necessary to detect this difference with only 1 chance in 10 of making a Type II error?
9. The credit manager for an oil company claims that the average balance on statements mailed to credit-card holders is at least \$32. To check this claim, an auditor takes a sample of 64 statements and finds that the average amount owed is \$30 with a standard deviation of \$12. On the basis of the sample evidence, what can we say about the credit manager's claim?
10. An auditor for another oil company takes a sample of 36 credit-card statements. He obtains a mean balance of \$34 and a standard deviation of

- \$10. Is there a significant difference in the mean balance of credit-card statements between this company and that of Problem 9 above?
11. Observations are made on the time required to check out customers in a supermarket. For a sample of 36 customers, it takes Mary an average of 6 minutes with a standard deviation of 3 minutes. It takes Joan an average of 8 minutes per customer with a standard deviation of 5 minutes for a sample of 36 customers. Is the difference in average time between the girls significant at the 5 percent level? (Use a two-tailed test.)
12. A coffee company was testing two new types of jars for its brand of instant coffee. To conduct the test 200 stores were selected, and each type of jar was introduced to one half of the stores. Sales records were kept for each store. The sales of the new jars were expressed as a percent of previous monthly sales. For jar *A*, the average sales increase was 3 percentage points with a standard deviation of 20 percentage points. For jar *B*, the average sales increase was 8 percentage points with a standard deviation of 24 percentage points.
- Is there significant evidence that the average sales increase for jar *A* is greater than 0 percent?
 - Is there significant evidence that the average sales increase for jar *B* is greater than 0 percent?
 - Is there a significant difference between the sample means?
13. Suppose two brands of cigarettes are tested for burning time with the purpose of deciding whether one brand is longer burning than the other. One hundred cigarettes of brand No. 1 are burned under test conditions, and the length of burning time is noted; 144 cigarettes of brand No. 2 are similarly tested for the length of burning time. The following means and standard deviations result (the subscripts referring to the brand number):

<i>Cigarette No. 1</i>	<i>Cigarette No. 2</i>
$n_1 = 100$	$n_2 = 144$
$\bar{X}_1 = 9.36$ minutes	$\bar{X}_2 = 9.00$ minutes
$s_1 = 0.83$ minutes	$s_2 = 1.20$ minutes

Estimate the difference in the mean burning time between the two brands and determine a 95 percent confidence interval for this difference.

14. The loan department of a certain bank specializes in loans to small businesses. For these loans, it is important to have an accurate evaluation of the financial standing of the business. To make this evaluation, a credit officer reviews the financial statements and application forms, and even interviews the applicant if desired and forms an opinion of the applicant's credit rating. This is expressed as an integer between 0 and 9, 9 being an excellent rating and 0 being the rating of a very poor credit risk.
- The management of the bank wished to be sure that the two credit officers, Green and Gray, were using the same standards in giving credit ratings. Accordingly, 30 applicants were selected at random, and Green and Gray were asked to make an independent evaluation. The results are shown below:

Application Number	Green Evaluation X_1	Gray Evaluation X_2	Difference d
1	8	7	1
2	5	3	2
3	6	7	-1
4	9	9	0
5	1	2	-1
6	4	2	2
7	5	5	0
8	8	6	2
9	7	4	3
10	5	6	-1
11	2	1	1
12	2	2	0
13	1	0	1
14	6	7	-1
15	5	4	1
16	3	3	0
17	6	6	0
18	6	5	1
19	4	5	-1
20	3	1	2
21	6	6	0
22	5	4	1
23	4	4	0
24	5	5	0
25	4	3	1
26	3	5	-2
27	4	3	1
28	8	9	-1
29	8	5	3
30	4	3	1
Total	147	132	+15
Mean	4.90	4.40	0.5
Sum of Squares	849	726	53

Management realized that there would be differences in the evaluation of individual applicants but wanted the credit officers to give the same average evaluation.

- a) Using the evaluations of the 30 applicants by Green and Gray as separate samples, test the hypothesis that there is no difference in their evaluations, on the average. Is the observed difference significant?
 - b) The fourth column in the above table shows the difference d between the evaluation of Green and Gray. Using this set of 30 observations as one sample, test the hypothesis that the mean of the difference d is equal to zero. Is the observed difference significant?
 - c) Compare the two methods of *a* and *b* for evaluating the differences between means. Why is the second more efficient than the first?
15. Refer to Problem 6 in Chapter 6. Is the observed difference in the averages of the two types of lamps significant?

SELECTED READINGS

Selected readings for this chapter appear in the list on page 314.

13. INFERENCES INVOLVING SMALL SAMPLES AND PROPORTIONS

IN THE TWO PREVIOUS CHAPTERS, the discussion of statistical inference has been based upon two assumptions: (1) a large sample was taken and (2) the sample statistic of interest was the sample mean, used as an estimate of the population mean. In this chapter, the discussion will concentrate on specific cases not covered by the two points above. In particular, the question of how to deal with small samples will be treated; and the sample proportion will be employed to make inferences about the population proportion.

SMALL SAMPLES

The assumption of large samples, which has been made up to this time, was necessary to insure (1) that the sampling distribution of the sample mean was approximately normal and (2) that there was little error introduced by estimating the population standard deviation σ by the sample standard deviation s . Because of these properties, large sample estimation is quite generally applicable, making possible statistical inferences without any specific assumption about the shape of the distribution from which the sample was drawn. But in certain situations it is not possible or economical to obtain a large sample. Does this mean that statistical probability statements cannot be made in these situations? The answer to this question is a strong "no," together with the qualification that additional assumptions or other methods are necessary. One method of dealing with small samples can be used when *the population distribution from which the sample is drawn is normal*, or approximately normal. There are two cases, depending on whether σ is known.

Case A: Sampling from a Normal Population, σ Known. The central limit theorem discussed in Chapter 11 states that means of large samples are approximately normally distributed. The same is true for small samples, provided the population from which the sample is drawn is normal (i.e., means of samples, both large and small, from normal populations are normally distributed). And if the standard deviation σ is known, the analysis can proceed exactly as in the previous two chapters. The standard error of the sample mean is, as before, $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. And confidence intervals for the population mean, as well as tests of hypothesis, can be formulated in the same way as before.

Case B: Sampling from a Normal Population, σ Unknown. When the population standard deviation σ is not known, it must be estimated from the data in the small sample. To handle the sampling error in both the same mean \bar{X} and the sample standard deviation s , a new sampling distribution must be introduced.

This symmetric but nonnormal distribution is called the t distribution. The ratio t (like the standard normal deviate u) is defined as the deviation of the sample mean from the population mean expressed in standard error units. That is,

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

where $s_{\bar{X}}$, the standard error of the mean, is computed from s , the standard deviation of a sample, by the formula $s_{\bar{X}} = s/\sqrt{n}$.

The sampling distribution of t differs for each size of sample. There is one t distribution for samples of size 10, another for size 11, and so on. Hence, the values of t corresponding to the 5 and 1 percent probability levels are not 1.96 and 2.58, as in the normal curve, but depend on the sample size, as shown in Table 13-1.

Table 13-1
VALUE OF t AT 5 AND 1 PERCENT PROBABILITY LEVELS

Degrees of Freedom	0.05	0.01
10	2.228	3.169
20	2.086	2.845
30	2.042	2.750
∞	1.960	2.576

Table 13-1 is abstracted from the more detailed t table in Appendix J. In this table the first column lists the "degrees of freedom" rather than sample size; that is, $n - 1$ instead of n in the examples used thus

far.¹ Since this column goes up to 30, we can define a small sample, for the purpose of using this table, as one in which n is 31 or less. The t distribution looks more and more like the normal distribution as n increases in size, so the t values approach the corresponding values for the normal distribution. These are listed in the last row of the table. The probabilities in the heading of the table refer to the sum of the two-tailed areas under the curve that lie outside the points $\pm t$. The values of t are listed in the body of the table. For a single-tailed area, divide the probability by 2.

As an example, for a sample of size 8, enter the row $n - 1 = 7$; then 5 percent of the area under the curve falls in the two tails outside the interval $t = \pm 2.365$. That is, $2\frac{1}{2}$ percent of the area falls in each tail, and 95 percent of the area falls within the interval $t = \pm 2.365$. A t value of 2.365 therefore should be used in setting up a 95 percent confidence interval for the mean when the sample size is 8.

Confidence Intervals

As an example, a manufacturer wishes to estimate the average weight of a large shipment of 20-gauge uncoated steel sheets received from a supplier. The estimate is to be expressed as a 95 percent confidence interval centered on a sample mean. He selects 8 pieces at random, and finds that the sample mean is 148.4 pounds per hundred square feet, while the standard deviation is 2.07 pounds. The standard error of the mean is then

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2.07}{\sqrt{8}} = 0.73 \text{ pounds}$$

To find the 95 percent confidence interval, he finds $t = 2.365$ in the table as described above. The confidence interval is then

$$\bar{X} \pm t \cdot s_{\bar{x}} = 148.4 \pm 2.365(0.73) = 148.4 \pm 1.7 \text{ pounds}$$

He can then state that the average weight of the whole shipment lies between 146.7 and 150.1 pounds, with a 95 percent chance of being correct.

Testing Hypotheses

Alternatively, the manufacturer in the foregoing problem might wish to test whether the mean weight of the sample of steel sheets (148.4

¹Gauss showed that "the number of observations is to be decreased by the number of unknowns estimated from the data, to serve as divisor in estimating the standard error." Here we use $n - 1$ because one degree of freedom is lost when the standard deviation is computed from a sample. When all deviations from the arithmetic mean but one are determined, the last one is also determined.

pounds) differs significantly from the specification of 150 pounds called for in his purchase order. So he computes the deviation of the sample mean from this hypothetical mean in units of the estimated standard error (0.73 pounds) as follows:

$$\begin{aligned}t &= \frac{\bar{X} - \mu_h}{s_{\bar{X}}} \\&= \frac{148.4 - 150}{0.73} \\&= -2.19\end{aligned}$$

Appendix J shows that for 7 degrees of freedom the 5 percent point of t is ± 2.365 , as noted above. Hence, the mean weight of 148.4 pounds does not differ significantly from the specified mean weight of 150 pounds at the 5 percent level of significance. If the absolute value of t had exceeded 2.365, the difference would have been considered significant at the 5 percent level.

The t test can be applied similarly to determine whether the *difference* between the means of two small samples is significant. This procedure will not be illustrated here.

In order to make inferences about the means of small samples when the population distribution is normal, then, we proceed as with large samples, except for using the t value in Appendix J in place of the corresponding normal value in Appendix D. When the population distribution is markedly nonnormal (especially when it is very skewed), other methods must be employed. Either techniques specific to the population being sampled or *nonparametric* methods,² which do not depend upon any particular distribution, may be used in these cases. Such techniques are treated in advanced statistical texts.

PROPORTIONS

The foregoing discussion of statistical inference has been applied to the arithmetic mean. This is an important measure of any variable. It should be noted, however, that many different statistical measures can be submitted to a similar type of statistical inference—medians, standard deviations, and so on. The three essential tools in such analysis are (1) the designated measure as found within the sample, (2) the standard error of the measure involved, and (3) the sampling distribution of the measure.

² For a discussion of some nonparametric methods, see Chapter 17 of the Dixon and Massey reference listed at the end of this chapter.

In this section we apply the principles of statistical inference to the *proportion*. As noted earlier, a proportion represents an *attribute* of a population rather than the average value of a *variable*. This might be the proportion of defective pieces in a lot of bolts produced, the proportion of customers that plan to buy a color television set, and so on.

It was pointed out in Chapter 5 that a proportion may be considered a special case of the arithmetic mean in which all the values are ones or zeroes. Our discussion about the sampling distribution of means thus applies for the most part to proportions also. In particular, the sample proportion is an unbiased estimate of the population proportion. That is, if all possible random samples were drawn from a population, the mean of the sample proportions, or the "expected value," would equal the population proportion. We will use the symbols p_s and p to denote the proportion of items in the sample and population, respectively, that have a given characteristic. Similarly, q_s and q denote the proportion of items that do *not* have that characteristic. Hence, $q_s = 1 - p_s$ and $q = 1 - p$.

The Binomial versus the Normal Distribution

The sampling distribution of a proportion (like that of the mean) is the distribution of its values that could be obtained from all possible random samples of size n taken from a population. Sample proportions follow the binomial distribution,³ though for larger samples (say, when np and nq are above 5) the normal approximation can be used instead, as described in Chapter 8.

We can set up confidence intervals and test hypotheses by use of a binomial table, such as Appendix F or G for samples up to 25 in size. For example, suppose we wish to test the hypothesis that $p \leq 0.20$ on the basis of a sample of 10 items, with a critical probability of 5 percent one-tailed. The sample result may produce 0, 1, 2, 3, etc. successes or the equivalent sample proportions of 0, 0.10, 0.20, 0.30, etc. From Appendixes F and G, we see that the probability of 0 or more successes, 1 or more successes, etc., up to 4 or more successes is in each case, more than 0.05, and only the probability of 5 or more successes is less than 0.05; that is, 0.033. Hence, the hypothesis can be rejected at the 5 percent level only if 5 or more successes (equivalent to a sample proportion of 50 percent or more) occur in the 10 sampled items.

However, statistical inference based on the binomial distribution

³ This is true assuming a very large population, or sampling with replacement. The reader is advised to review Chapter 8 on the binomial distribution and its normal approximation before proceeding.

involves complex technical difficulties, such as those arising from the discreteness of the distribution and the asymmetry of confidence intervals. Further, it is difficult to make a valid inference based on a small sample alone (when the normal approximation cannot be used), without also considering prior information. We will show how to combine prior information and binomially distributed sample data for decision-making in Chapter 15. In the present chapter, therefore, we will restrict the discussion to large samples (where np and nq are over 5), so that a nearly normal distribution can be assumed. The analysis is thereby simplified, and the concepts developed for the mean in Chapters 11 and 12 can be carried over and applied directly to the proportion.

The Standard Error of a Proportion

The standard error of a proportion is the standard deviation of the p_s 's in all samples that might be drawn from a population. As in the case of the mean, the standard error of a proportion equals the standard deviation of the population divided by the square root of the sample size. In the case of the proportion, however, the standard deviation of the population is $\sigma = \sqrt{pq}$. Hence the standard error of a proportion is

$$\sigma_{p_s} = \sqrt{\frac{pq}{n}}$$

For example, if $n = 100$ and $p = 0.20$,

$$\sigma_{p_s} = \sqrt{\frac{0.20 \times 0.80}{100}} = \frac{0.40}{10} = 0.04 \quad \text{or 4 percent}$$

As in the case of the mean, the standard error of a proportion depends on the absolute size of the sample n , rather than on its relation to the size of population n/N .⁴

⁴ If the sample makes up a large part of the population, however, the same finite population correction applies as in the case of the mean. The formula is then

$$\sigma_{p_s} = \sqrt{\frac{pq}{n}} \sqrt{1 - \frac{n}{N}}$$

Thus, if the whole lot or population had a size of only $N = 500$ in the above example, we would have

$$\begin{aligned} \sigma_{p_s} &= \sqrt{\frac{0.20 \times 0.80}{100}} \sqrt{1 - \frac{100}{500}} \\ &= 0.04 \times 0.9 = 0.036 \end{aligned}$$

The Confidence Interval for a Proportion

Suppose that the management of a large grocery chain is interested in estimating what proportion of its customers would prefer a self-service display of prepackaged meat to a meat counter serviced by a butcher. The market research department is assigned to make a study leading to such an estimate.

A random sample of 400 customers is taken, and it turns out that 220, or 55 percent, are in favor of the self-service display. It is extremely unlikely that the population constituting *all* customers would divide in preference exactly in this proportion. How, then, do we estimate the interval in which the true proportion falls with, say, a 95 percent degree of confidence? The analytical principles are the same as those used in constructing confidence intervals for the arithmetic mean. Only the measures are altered to fit the present case.

The standard error of a proportion, as we saw a moment ago, ideally requires the population value of p for its calculation. This we do not know, or we would not be faced with the problem of estimating the interval within which it falls. The common practice is to assume that p has the value of p_s found in the sample and to make the substitution accordingly. Hence, the estimated standard error for the sample proportion is⁵

$$\begin{aligned} s_{p_s} &= \sqrt{\frac{p_s q_s}{n}} \\ &= \sqrt{\frac{0.55 \times 0.45}{400}} \\ &= 0.0249 \text{ (rounded to 0.025)} \end{aligned}$$

Using the normal distribution, the 95 percent confidence interval is $p_s \pm 1.96s_p$, or about two standard errors on each side of 0.55. Therefore, we are 95 percent confident that the true proportion of customers favoring self-service meat counters lies somewhere between 50 and 60 percent.

As in the case of the arithmetic mean, and for the same general reasons, we could construct intervals of varying degrees of confidence, based upon appropriate multiples of the standard error of the proportion laid off around the value for p_s observed in the sample.

⁵ The formula shown is the one almost universally used, although it is biased. An unbiased estimator would have $n - 1$ in the denominator instead of n . However, for large samples, the difference is trivial. See W. Cochran, *Sampling Techniques*, 2d ed. (New York: John Wiley, 1963), p. 33.

Size of Sample. The *size* of a simple random sample needed to reduce the standard error to any desired level can be computed from the above formula in the same way as with the mean. Suppose we wish to determine the proportion of customers preferring self-service with a sample standard error of only 0.02, or two percentage points. This corresponds to 95 percent confidence limits of $p_s \pm 1.96(0.02)$ or $p_s \pm 0.04$. From the trial survey cited above, p is tentatively 0.55. Then we solve for n in the equation $s_p = \sqrt{(p_s q_s)/n}$, as follows:

$$0.02 = \sqrt{\frac{0.55 \times 0.45}{n}}$$

$$\text{Transposing, } \sqrt{n} = \frac{\sqrt{0.55 \times 0.45}}{0.02} = \frac{0.4975}{0.02} = 24.9$$

$$\text{Squaring, } n = 620$$

It is necessary to sample about 620 customers (or 220 in addition to those already sampled), therefore, in order to obtain a value of p_s that has a standard error of only 0.02.

The Test of a Hypothesis for a Proportion

Let us suppose that the preceding problem has come up in a somewhat different way—and for purposes of exposition assume that we know nothing of the calculations made in the foregoing section.

Assume that a nationwide survey by a grocery trade association had suggested that customers of chain stores were equally divided in their preference between self-service meat counters and counters serviced by butchers. The management of a regional chain is somewhat impressed by this finding, but it recognizes that regional differences can exist. Management has decided that it will replace butcher-serviced counters if it can get compelling evidence that its particular group of customers favors self-service in a proportion greater than one half.

Now, in this case the nationwide survey has suggested the hypothesis that the true proportion is 0.50, and only if this is refuted by regional evidence will management decide otherwise. Further, management is interested only in the alternative hypothesis that the true proportion is *greater than* 0.50; therefore a one-tailed test is the appropriate one.

Let us assume that a random sample of 400 customers is drawn. From the hypothesis that the true population proportion is 0.50 (i.e., $p_h = 0.50$), we proceed to calculate the standard error of a sample proportion which would correspond to that hypothesis, namely,

$$\begin{aligned}
 \sigma_{p_s} &= \sqrt{\frac{p_h q_h}{n}} \\
 &= \sqrt{\frac{.50 \times .50}{400}} \\
 &= .025
 \end{aligned}$$

Suppose that the proportion of customers favoring self-service in the sample turns out to be 0.55; then the difference between the sample proportion (p_s) and the hypothetical proportion (p_h) is 0.05. In terms of multiples of the standard error, this is

$$\frac{p_s - p_h}{\sigma_{p_s}} = \frac{0.55 - 0.50}{0.025} = \frac{0.05}{0.025} = 2$$

Only 2.3 percent of the area under a normal curve falls *above* 50 percent by more than two standard errors in that one-tailed direction (see Appendix D). Hence, the probability is only 2.3 percent that such a large proportion could occur by chance if the true proportion were no greater than 0.50. We should have to make our decision on the grounds discussed earlier. But the probability of 2.3 that chance alone could have created this evidence is surely a low probability. And a conclusion that the true population proportion is greater than 0.50 is strongly indicated.

The Test of a Difference between Two Proportions

Suppose that a manufacturer of farm implements is interested in whether farmers in state No. 1 differ significantly from farmers in state No. 2 with respect to the proportion preferring the make of tractor which he sells. He takes separately a random sample of 100 farmers in each state and finds that the proportion preferring his make is 0.40 in state No. 1 and 0.30 in state No. 2. Should this difference in sample proportions be taken as signifying a difference in the true proportions?

The line of statistical reasoning by which this question is answered is already familiar from earlier discussions. Only the new, appropriate measures need to be introduced. The sampling distribution of ($p_{s_1} - p_{s_2}$) may be taken to be fairly normal in large samples because of considerations discussed in the last section.

The *standard error of a difference* between two independent sample proportions p_{s_1} and p_{s_2} is

$$\sigma_{p_{s_1} - p_{s_2}} = \sqrt{\sigma_{p_{s_1}}^2 + \sigma_{p_{s_2}}^2}$$

Since the symbolism is going to be a little complicated, it will be more convenient to write this in squared form, which is known as the *sampling variance* of the difference between two proportions. Hence,

$$\sigma_{p_{s_1}-p_{s_2}}^2 = \sigma_{p_{s_1}}^2 + \sigma_{p_{s_2}}^2$$

That is, the sampling variance of the difference between two independent proportions is the sum of their sampling variances.⁶

Since $\sigma_{p_s}^2 = pq/n$ in each case, the above formula may be written

$$\sigma_{p_{s_1}-p_{s_2}}^2 = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$$

in which the subscripts 1 and 2 refer to the two states, respectively.

Now, in the present case, we would set up and test the *null hypothesis* that there is *no* difference in the true population proportions involved. Our hypothesis states that $p_1 = p_2$; hence, the observed difference between the sample proportions p_{s_1} and p_{s_2} is caused by sampling errors.

Since we do not know p_1 and p_2 , the best estimate of their common value is the weighted mean of the sample proportions (using the sample sizes as weights). This is most easily accomplished by adding the *number* of farmers preferring the tractor in both samples and dividing this total by the total number of farmers. There are 70 farmers preferring the tractor (40 from state No. 1 and 30 from state No. 2) out of 200 farmers sampled, and so the weighted mean proportion is $\bar{p} = 70/200 = 0.35$.

The sample variance then is

$$\begin{aligned}\sigma_{p_{s_1}-p_{s_2}}^2 &= \frac{\bar{p}q}{n_1} + \frac{\bar{p}q}{n_2} \\ &= \frac{0.35 \times 0.65}{100} + \frac{0.35 \times 0.65}{100} \\ &= 0.00455\end{aligned}$$

To find the standard error of the difference we extract the square root, which gives

$$\sigma_{p_{s_1}-p_{s_2}} = 0.0675$$

In the way now familiar, we express the observed difference of the sample results from the null hypothesis as a ratio to the standard error of such differences. Since the null hypothesis assumes the true difference to be zero, the calculation which we want amounts to

⁶ As a graphic solution or check, lay off $\sigma_{p_{s_1}}$ and $\sigma_{p_{s_2}}$ as the sides of a right triangle; then $\sigma_{p_{s_1}-p_{s_2}}$ is the hypotenuse.

$$\frac{p_{s_1} - p_{s_2}}{\sigma_{p_{s_1} - p_{s_2}}} = \frac{0.40 - 0.30}{0.0675} = 1.48$$

so that the observed difference deviates from the null hypothesis by 1.48 standard errors.

Consultation of Appendix D shows that deviations of this size, regardless of sign, from a true value of zero, are expected to occur by chance alone in 14 percent of all possible samples. In other words, the probability is about 14 percent that this big a spread could occur by chance alone, were the null hypothesis true. This is not significant at the 5 or 10 percent level. Therefore, based on the available evidence, we would probably "accept the null hypothesis" and attribute the sample results to mere chance. We do not have sufficient evidence to reject the null hypothesis, that is, to conclude that there is a real difference between the two states sampled. This does not prove that $p_1 = p_2$; the evidence is inconclusive. The manufacturer should consider increasing the size of the samples, so that for any given critical probability chosen the overall likelihood of committing an error of inference would be reduced.

SUMMARY

Small Samples. If small samples are drawn at random from a *normal* population, and the parameter σ is known, then the sample means also follow a normal distribution, and we can make statistical inferences exactly as done in Chapters 11 and 12.

If small samples are drawn from a normal population and σ is *not* known, however, the sampling errors in \bar{X} and s cause the means to follow a *t distribution*, which differs more and more from the normal distribution as the sample size become smaller. We should then look up $t = (\bar{X} - \mu)/s_{\bar{X}}$ in Appendix J (if $n \leq 31$) to find the appropriate multiple of the standard error for use in setting up confidence limits or testing hypotheses.

Finally, if small samples are drawn from a population that is markedly *not* normal, neither of the above methods applies, and more advanced techniques must be used.

Proportions. Inferences may be made about sample proportions in much the same way as with means. In fact, a proportion may be considered a special case of a mean in which the attributes, such as defectives and nondefectives, are valued 1 and 0, respectively, and averaged to find the percent defective.

The standard error of a proportion is $\sigma_{p_s} = \sqrt{(pq)/n}$, where p is the population proportion and $q = 1 - p$. This is estimated as $s_{p_s} = \sqrt{(p_s q_s)/n}$ when sample values are used.

The sampling distribution of p_s follows a binomial distribution, but for large samples (say, when np and nq are greater than 5) the distribution is approximately normal, so we assume normality here both because it is valid for most practical problems and because it is simpler than using the binomial distribution.

A 95 percent *confidence interval* may then be laid out around the sample proportion (i.e., $p_s \pm 1.96s_{p_s}$) to include p , the population proportion, with a 95 percent chance of being correct. Other degrees of confidence are handled similarly.

The *size* of sample needed to reduce the standard error s_{p_s} to any desired value can be obtained by solving for n in the formula $s_{p_s} = \sqrt{(pq)/n}$, using an estimated value of p .

Tests of hypotheses may be applied to proportions by computing the standard error, based on the hypothesized proportion p_h . Then the deviation of the sample proportion from this value ($p_s - p_h$) is divided by the standard error to determine whether it is large enough to be significant. Thus, if the standardized deviation is 1.96 or more (in a two-tailed test), it is significant at the 5 percent level of confidence, and so on (Appendix D).

We can also test whether the *difference between two proportions* ($p_{s_1} - p_{s_2}$) is significant by dividing the difference by its standard error, where $s_{p_{s_1} - p_{s_2}}^2 = s_{p_{s_1}}^2 + s_{p_{s_2}}^2$. If this standardized difference is 1.96 or more, it is significant at the 5 percent level etc., just as above. When we test the *null hypothesis* that there is no difference between p_1 and p_2 , we use the average value of the sample proportions, weighted by the size of the two samples, to compute the standard error of the difference.

PROBLEMS

1. Explain:

- a) Why the means of large samples follow the normal distribution while the means of small samples may deviate significantly from normality.
- b) Why, when taking a small sample from a normal population, the normal distribution can be used for statistical inference if σ is known while the t distribution must be employed if σ is not known.

2. Explain:

- a) The concept of the proportion as a special case of the mean.
- b) The relation between the distribution of proportions and the normal distribution.

- c) A 90 percent confidence interval for a proportion.
 - d) How to test a hypothesis that a sample proportion 0.45 is significantly less than 0.50.
 - e) The null hypothesis for the difference between two sample proportions.
 - 3. Management is interested in the average wait for a customer at a checkout counter during certain peak periods in a supermarket. A sample of 16 customers is taken at random, and their waiting times are noted. The mean waiting time was 7 minutes with a standard deviation of 3 minutes. Can we conclude (with 95 percent confidence) that the mean waiting time was *not* less than 5 minutes? (Assume that the population to be sampled is normal.)
 - 4. A random sample of 25 is drawn from the records of daily output of a large group of employees in order to estimate the population mean. The sample shows a mean of 136 units and a standard deviation of 24 units. (Daily output is normally distributed.)
 - a) Calculate a 98 percent confidence interval for the mean output of all employees.
 - b) Does the mean output of 136 units differ significantly from the standard output of 144 units set by management? Explain.
 - 5. A survey of consumer buying plans reports that 10 percent of a sample of 2,500 families plan to buy a new refrigerator during the next year. Assume that an unbiased simple random sample was used. Set up a 99 percent confidence interval to estimate total refrigerator sales for the whole population of 50 million families. Interpret this forecast.
 - 6. The consumer research division of an automobile manufacturing firm has a budget of \$3,000 for a survey to determine the proportion of consumers who prefer a new design for the radiator grill. The estimate should be correct to within 5 percentage points, with a 95 percent confidence coefficient. Assume a simple random sample. Cost of the survey is \$1,000 for overhead plus \$5 an interview.

Can this proportion be estimated with the required precision for \$3,000, assuming $p = 0.50$? Explain.
 - 7. A television distributor finds that about 22 percent of the potential customers who enter his store buy a television set. Moving to another city, he wishes to estimate this percent for the new location within ± 4 percent, at the 90 percent confidence level. How many observations should he take?
 - 8. The median life of a certain electronic tube is claimed by the manufacturer to be 600 hours. You draw a random sample of 100 from a shipment of these tubes and find that only 23 last over 600 hours. Do you believe the manufacturers' claim? Why? (Hint: 50 percent of the values exceed the median.)
 - 9. After finding that 23 out of 100 electronic tubes from manufacturer No. 1 outlast 600 hours, you order a shipment of similar tubes from manufacturer
-

No. 2 and find that 52 out of a random sample of 200 outlast 600 hours. Is there a significant difference in the durability of the two manufacturers' tubes? Explain.

10. If, in a sample of 600 economics students drawn from schools throughout the country, 360 are sons of businessmen, what is the 90 percent confidence interval for the proportion of *all* economics students who are sons of businessmen?
11. You wish to make a market survey to estimate the proportion of housewives who prefer your new product to competitors' products. You would like the error in estimating the proportion to be no greater than 4 percent points, with a confidence coefficient of 95.45 percent. The sales department offers a preliminary guess that about 20 percent of housewives might prefer your product. If the survey costs \$500 to set up, and \$5 an interview, about how much should the whole survey cost?
12. A production supervisor wished to estimate the percent of time a certain machine was idle because of breakdowns, delays, etc. Since it would be difficult to keep accurate records, a sampling procedure was instituted. Accordingly, the status of the machine was checked by the supervisor over a period of four weeks at random times (i.e., the times were selected in advance, using a table of random numbers). This procedure is known as *work sampling*. A total of 300 checks were made on the machine, and in 24 instances the machine was idle.
 - a) Estimate the percent of idle time on the machine and calculate a 90 percent confidence interval about the estimate.
 - b) Determine if the percent of idle time is significantly less than 10 percent?
13. The Alvin Chemical company is contemplating adding some petroleum storage tanks at its distribution center in Chicago. It is common practice in this company to obtain several estimates from its own engineers of the cost of such capital expenditures. The average of these estimates is then used as the expected expenditure figure in capital budget planning.
For the storage tanks in Chicago, five estimates were obtained:

<i>Estimator</i>	<i>Estimate</i> (Millions of Dollars)
Pearson	\$ 9
Neyman	14
Fisher	8
Wald	9
Hotelling	10

Noting the diversity in the estimates, Mr. Alvin, the president, wonders if it would not be possible to put some outside limits (say with 95 percent confidence) as maximum and minimum estimated expenditures.

- a) Provide Mr. Alvin with such an interval estimate.
- b) What assumptions is it necessary to make to give this estimate? Discuss the validity of these assumptions.

14. The following are data obtained by the management of a department store in a study of delinquent time payment accounts: In a sample of 600 time-payment accounts opened by individuals who had resided in the community for more than 5 years, 58 had become delinquent at one time or another. In a sample of 400 time-payment accounts for individuals who had resided in the community for less than 5 years, 26 had become delinquent.
- Is the difference between the two significant at the 5 percent level?
 - What is a possible fallacy in interpreting this difference, whether significant or not?

15. The market research department of the Bodhauser Beer Company conducted a taste test to determine if consumers could distinguish Bodhauser Beer from its chief competitor, Schultz. Accordingly, 200 beer drinkers were selected, given unmarked samples of both beers, and told to state a preference.

Because it was feared that the order in which the different beers were presented to the test group might affect their preference, the group was broken into two parts; half the group were given Bodhauser before Schultz, and the other half were given Schultz before Bodhauser.

The results are shown in the table below:

	Group 1 Bodhauser Before Schultz	Group 2 Schultz Before Bodhauser
Number in group	100	100
Number preferring Bodhauser	54	58

- Ignoring the order in which the beer was presented (i.e., lumping both groups together), was there significant evidence that either beer was preferred over the other (i.e., Schultz over Bodhauser or vice versa)?
- Were the initial fears that the order might affect the preference substantiated? That is, is there evidence from the experimental data that the two sampled groups differed?

SELECTED READINGS

DIXON, W. J., and MASSEY, F. J. *Introduction to Statistical Analysis*, 2d ed. New York: McGraw-Hill, 1957.

An excellent reference source on the use of statistical inference in a variety of situations. Chapters 7, 8, and 9 discuss statistical inference, estimation, and tests of hypotheses. Chapter 14 discusses Type II error in detail.

GUENTHER, WILLIAM C. *Concepts of Statistical Inference*. New York: McGraw-Hill, 1965.

An extended treatment of inference at the elementary level.

HOEL, PAUL G. *Introduction to Mathematical Statistics*, 3d ed. New York: John Wiley, 1962.

Presents the mathematical foundations of statistical inference at an intermediate level for those with a background in calculus.

KEEPING, E. S. *Introduction to Statistical Inference*. Princeton, New Jersey: Van Nostrand, 1962.

Emphasizes the inferential and decision-making aspects of statistics at the intermediate mathematical level.

MCCARTHY, PHILIP J. *Introduction to Statistical Reasoning*. New York: McGraw-Hill, 1957.

Treats inference in the form of conceptual models, with emphasis on the binomial distribution.

RICHMOND, SAMUEL B. *Statistical Analysis*. 2d ed. New York: Ronald Press, 1964.

Chapters 6 to 8 provide a rigorous treatment of statistical inference.

SCHLAIFER, ROBERT. *Introduction to Statistics for Business Decisions*. New York: McGraw-Hill, 1961.

Chapters 10 and 11 present a treatment of the classical theory of statistical inference from a Bayesian decision-theory approach.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*. New York: The Free Press, 1956.

Part III treats a wide variety of topics in inference, with many examples.

14. SAMPLE SURVEY METHODS

MUCH OF THE MATERIAL that we have studied has been concerned with the interpretation and evaluation of sample information. The emphasis has been primarily on simple random samples. In actual practice, simple random samples are often impossible to obtain or prohibitively expensive. In this chapter we examine some different methods of selecting samples. Some of these methods will be more efficient than simple random sampling; others can be used where simple random sampling would be impossible; others are less costly than simple random sampling.

There are two broad classes of methods of selecting samples: (1) *probability sampling*, including simple random sampling, systematic selection, stratified random sampling, ratio estimation, and cluster sampling¹ and (2) *nonprobability sampling*, including quota sampling and judgment sampling. These are discussed below.

PROBABILITY SAMPLING

Probability sampling includes all methods of sampling in which the sampling units are selected according to the laws of chance so that the probability of being included is known (and not zero) for each member of the population. "Selected according to the laws of chance" means using some chance device such as a table of random numbers rather than personal judgment to choose the items sampled. The "probability of being included" may be *equal* for all units in the population (as in simple random sampling), or it may be, say, "probability proportional

¹ The application of probability sampling to statistical quality control is discussed in Chapter 25. This includes sequential sampling plans for the control of a process, as well as acceptance sampling, which is used to determine whether to accept or reject industrial products.

to size" (e.g., a company with 2 million sales having twice the probability of being selected as one with 1 million sales). In any case, however, the probability must be *known*, and hence the population itself must be identifiable.

In probability samples one can estimate objectively the precision of the sample results or compare the precision of different types of samples. The precision of probability samples increases (i.e., the sampling error decreases) as the size of the sample increases, whereas errors of judgment persist in larger nonprobability samples. Hence, probability sampling is generally used, wherever feasible, in large-scale surveys.

Simple Random Sampling

Simple random sampling has been used in all our discussions of sampling in Chapters 11 through 13. Simple random sampling means that each possible sample of a given size in the population has an equal chance of being selected. It is nothing new. This section is thus offered as a review.

1. **Estimation of Mean and Variance.** $\bar{X} = \Sigma X/n$ is an unbiased estimate of μ , the population mean, and $s^2 = \Sigma (X - \bar{X})^2 / (n - 1)$ is an unbiased estimate of σ^2 , the population variance.

2. **Sampling Error.** The estimate of the standard error of the sample mean is

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

3. **Finite Population Correction.** When sampling without replacement from finite populations, the finite population correction (fpc) is included in the standard error of estimate. This factor can be ignored as negligible if the sample comprises less than 5 percent of the population. The standard error of the mean for finite populations is

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

where n is the sample size and N is the population size.

4. **Population Totals.** Estimates of a population total and the standard error associated with the estimate of the total may be made simply by multiplying the sample mean \bar{X} and the standard error of the mean $s_{\bar{X}}$ by the number of items in the population N . Thus,

$$\begin{aligned} \text{Estimate of Population Total} &= T = N\bar{X} \\ \text{Estimated Standard Error of Population Total} &= s_T = Ns_{\bar{X}} \end{aligned}$$

5. Proportions. When we are sampling to estimate the proportion of the population having a certain characteristic, the sample proportion p_s is an unbiased estimate of the population proportion p . Then the estimate of standard error (where n/N is small) is

$$s_{p_s} = \sqrt{\frac{p_s q_s}{n}}$$

The estimated standard error for finite populations is

$$s_{p_s} = \sqrt{\frac{p_s q_s}{n}} \sqrt{1 - \frac{n}{N}}$$

Systematic Selection

A systematic sample is one in which every k th item (e.g., every tenth item) is selected in a list representing a population or a stratum (a relatively uniform segment) of the population. The number k is called the *sampling interval*. The first number is chosen at random from the first k items, as described below. Systematic selection ensures that the items sampled will be spaced evenly throughout the population.

For example, suppose you wish to take a systematic sample of six households from a block of 78 households. First, list and number the households. Then divide 6 into 78; this means that you should select every thirteenth house. Choose the first household at random from the numbers 1 through 13, using a table of random numbers. Say this is number 6. Now select every thirteenth house beginning with number 6—that is, 6, 19, 32, 45, 58, and 71—to complete the sample.

Systematic sampling is often equivalent in its results to random sampling, if the elements in the population occur in a random order. For example, in dealing cards in the game of bridge, each player has a systematic sample (every fourth card). If the cards are shuffled well before the deal, the hand is equivalent to a random sample. Where the elements in the population are considered in random order, the formulas used above for simple random sampling apply also to systematic sampling.

Systematic selection has an important advantage over simple random sampling if similar parts of the population tend to be grouped together, that is, if nearby elements resemble each other more than they resemble those at greater distances. For example, residents with similar incomes tend to be located close to one another. A systematic selection of a city's blocks, numbered in serpentine fashion as described below, would then

include more nearly the same proportion of each income group than a simple random sample.

Systematic selection should not be used, however, if there is some periodic variation in the population corresponding to the sampling interval. For example, in the case of sampling households in a block, if the block were laid out so that every eighth house were a large one on the corner, a systematic sample of every eighth house might include only large corner houses.

Systematic sampling has come into widespread use because it is easy to apply and it usually yields good results. For example, in the 1960 Census of Population every fourth person was asked several supplementary questions on housing. The cost of collecting and compiling information for this 25 percent sample was small compared with that of a complete enumeration or of an independent 25 percent sample survey. At the same time, the reliability of the information was sufficient for almost any purpose.

Stratified Sampling

If a population is made up of fairly uniform parts or strata, the precision of sample results can be improved by *stratification*. That is, the population is first broken down into strata, such that the elements within each stratum are more alike than the elements of the population as a whole. Then an assigned part of the sample is drawn from each stratum by random selection (or by one of the other methods to be described later). Stratification is therefore only one step in the complete sampling method; it is always used in conjunction with other procedures.

As indicated above, the strata should be defined so that the significant elements within a stratum are more uniform than they are for the population as a whole. For example, in a study of household incomes a city can be divided into high- and low-income areas so that income varies less within each area than it does in the city as a whole. Here, geographic location provides a useful basis for stratification. In this case, the average income of a stratified random sample generally will be closer to the true average for the whole population than would that of a simple random sample of the same size selected from the city as a whole without stratification. Stratified sampling is thus useful for reducing the sampling error. As an extreme example of how stratification reduces this error, consider the following. A factory has only two categories of workers, each category having only one wage rate. If we were to take a simple random sample of workers in the factory and measure wages, we

would have an estimate and some sampling error associated with the estimate. However, if we were able to group the workers by classification into two strata, we could then take a sample of only one worker for each stratum, and we would have no sampling error at all. We would know exactly the wages of all in the factory.

While the above example is artificial, it does illustrate the fact that by taking homogeneous groups, and sampling separately from each group, we can gain some accuracy in sampling. A second advantage of stratification is that it gives us separate estimates for parts of the population. This kind of information may be useful for management in planning advertising and so on.

Stratification should therefore be applied to *heterogeneous* populations, such as humans, since people can be divided into fairly uniform strata—by income, sex, age, or other criteria that affect the variable being studied (e.g., buying habits). Under these circumstances, stratification usually achieves greater precision for a given cost. On the other hand, stratification is unnecessary in *homogeneous* populations, as in measuring the diameter of ball bearings, where there are no discernible strata, such as differences in machine tools or operators, that affect the results.

Example. As an illustration of the use of stratified sampling, let us consider an application in the railroad industry.²

The bill for goods shipped (called a waybill) is usually paid to one railroad. However, the goods may have traveled over several different railroads while going from shipper to receiver. Each railroad over which the goods traveled is allocated a portion of the total revenue of the waybill. At one time, this was done by examining all waybills and allocating the revenue on each. A sampling procedure was considered to reduce the accounting cost of estimating the revenue allocation between railroads.

Table 14-1 shows the distribution of revenues of waybills terminating at a certain junction. Note that this distribution is extremely skewed, with a large number of waybills having small dollar amounts and a few having large amounts. It was decided to stratify the population into five groups—the same as those shown below. The waybills were accordingly sorted, and the number of waybills and total freight revenue in each group were ascertained. A systematic sample of each group was selected

² This example is adapted from C. West Churchman, "Applications of Sampling to LCL Revenue Divisions," in *Proceedings: Modern Statistical Methods for Business and Industry*, (Pittsburgh: Graduate School of Industrial Administration, Carnegie Institute of Technology, May, 1953).

Table 14-1

FREQUENCY DISTRIBUTION OF WAYBILLS

Waybill Revenue	Number of Waybills	Percent of Waybills	Total Revenue	Percent of Total Revenue
0 to \$ 4.99	3,047	56.0	\$ 8,868	15.5
\$ 5 to \$ 9.99	1,074	19.7	7,502	13.1
\$10 to \$19.99	645	11.8	8,934	15.6
\$20 to \$39.99	381	7.0	10,695	18.7
\$40 and over	298	5.5	21,245	37.1
Total	5,445	100.0	\$57,244	100.0

as shown in Table 14-2. Note how the proportion of each stratum sampled varies from 5 percent of Group 1 to 100 percent of Group 5. This is an efficient procedure for extremely skewed distributions such as we have here.

Using the percentages of revenue accruing to each railroad in each group (stratum), it is possible to estimate the percent of total revenue due each railroad.

Table 14-2

STRATIFIED SAMPLE OF WAYBILLS

Group	Waybill Revenue	Waybills Selected in Sample All Waybills Nos. Ending in	Approximate Percentage Sample
1	\$ 0 to \$ 4.99	02, 22, 42, 62, 82	5
2	\$ 5 to \$ 9.99	2	10
3	\$10 to \$19.99	2 and 4	20
4	\$20 to \$39.99	01 through 50	50
5	\$40 and over	All	100

Estimate of the Mean and Standard Error. Before introducing the estimation formula for stratified sampling, it is necessary to introduce some notation: Let M_i = the number of elements (items) in the i th stratum; N = total number of elements in the population = $\sum M_i$; m_i = the sample size in the i th stratum; \bar{Y}_i = the mean of the sampled elements in the i th stratum; s_i = the sample standard deviation in the i th stratum. Then

$$\text{Estimate of overall mean} = \bar{Y}_s = \sum w_i \bar{Y}_i$$

where w_i represents the weight of the i th stratum, computed as

$$w_i = \left(\frac{M_i}{N} \right)$$

$$\text{Standard error of overall mean} = s_{\bar{Y}_s} = \sqrt{\sum w_i^2 s_{\bar{Y}_i}^2}$$

where $s_{\bar{Y}_i}$ is the estimated standard error in each stratum. That is,

$$s_{\bar{Y}_i} = \frac{s_i}{\sqrt{m_i}} \sqrt{1 - \frac{m_i}{M_i}}$$

(The last term is the finite population correction—this can be ignored in any stratum in which the sample size m_i is less than 5 percent of the total number of elements in the stratum M_i .)

A few comments will help the understanding of these formulas. Note that the weight w_i is simply the fraction of the population in the i th stratum. The overall mean is simply a weighted average of the means in each stratum, using the relative numbers in each stratum as the weights. The standard error is weighted in a similar fashion.³

A simple example will help to clarify further the meaning of the formulas. Suppose we wish to estimate the mean annual income of a population, which we divide into two strata—a high and a low income group. The first stratum is composed of 1,000 members of which we sample 100. The second stratum contains 2,000 members of which we sample 500. These numbers are shown, together with the sampling results, in Table 14-3.

Table 14-3
STRATIFIED SAMPLE OF INCOMES

Stratum Number (i)	Number of Items in Stratum (M_i)	Number of Items in Sample (m_i)	Mean Income of Items in Sample (\bar{Y}_i)	Standard Deviation of Items in Sample (s_i)
1	1,000	100	\$10,000	\$1,000
2	2,000	500	5,000	500
Total	3,000 = N	600		

To estimate the average (\bar{Y}_s) for the total population we first have to determine the weights to attach to each stratum. These are:

$$\text{Weight for first stratum} = w_1 = \frac{1,000}{3,000} = \frac{1}{3}$$

$$\text{Weight for second stratum} = w_2 = \frac{2,000}{3,000} = \frac{2}{3}$$

³ Actually, the variance is weighted by w_i^2 .

That is, one third of the population items are in the first stratum, and two thirds are in the second stratum. Then the estimate of the population mean is

$$\bar{Y}_s = \sum w_i \bar{Y}_i = (1/3)(\$10,000) + (2/3)(\$5,000) = \$6,667$$

We next wish to calculate the standard error for this estimate. To do this we must first calculate the standard errors of the mean for each stratum:

$$s_{\bar{Y}_i} = \frac{s_i}{\sqrt{m_i}} \sqrt{1 - \frac{m_i}{M_i}}$$

That is,

$$s_{\bar{Y}_1} = \frac{1,000}{\sqrt{100}} \sqrt{1 - \frac{100}{1,000}} = \sqrt{9,000}$$

$$s_{\bar{Y}_2} = \frac{500}{\sqrt{500}} \sqrt{1 - \frac{500}{2,000}} = \sqrt{375}$$

And the standard error for the population mean is

$$s_{\bar{Y}_s} = \sqrt{\sum w_i^2 s_{\bar{Y}_i}^2} = \sqrt{(1/3)^2(9,000) + (2/3)^2(375)}$$

$$= \sqrt{1,167} = \$34$$

It can be demonstrated—though it is not done here—that a simple random sample of 600 items from this same population would have yielded a sampling error of about \$100. Hence, stratification was quite efficient in this example.

Allocation of the Sample to Strata: Proportional Allocation. In the example above, we arbitrarily established sample sizes of 100 and 500 in the two strata, respectively. Now, our knowledge of survey sampling procedures is of primary usefulness in designing surveys *beforehand* rather than *ex post facto*. Hence, the student may wonder at such an allocation of sample items between strata. Would it not have been better to have them more equally distributed? How large a sample should be taken in each stratum?

One simple answer to this is *proportional* allocation, that is, allocate items in the sample to the various strata in the same proportion as the total elements in the population.

As an illustration, suppose that the example given above represented a sample taken a year ago and that we are going to design a new sample. (Assume that the number of elements in each stratum and the standard deviations in each stratum remain the same.) Suppose that our new

sample will also be 600 items, but we are free to allocate these items between the two strata as we see fit.

Proportional allocation would mean that since there are $\frac{1}{3}$ of the items of the total population in the first stratum, then $\frac{1}{3}$ of the sample items should also be in that stratum. Thus, $m_1 = \frac{1}{3}$ of $600 = 200$. And since there are $\frac{2}{3}$ of the items in stratum 2, it should receive $\frac{2}{3}$ of the sample. That is, $m_2 = \frac{2}{3}$ of $600 = 400$. Proportional allocation is used if (1) the variability within the strata is approximately constant (i.e., the standard deviations within each of the strata— s_i —are about the same) or (2) little is known about the variability within the strata (hence, we may as well assume that they are about the same).

Proportional allocation has several advantages. It is the intuitively plausible or common-sense method of representing the different parts of the population (like the Supreme Court's proportional representation decree for state legislatures). In addition, it sometimes makes the formulas easier. For example, the estimate of the mean of the population is simply the mean of the sample—no weights are necessary.⁴

Allocation of the Sample to Strata: Optimum Allocation. If there is a considerable amount of variability within the strata, however (i.e., the standard deviations of the items in the strata—the s_i —are of different magnitudes), we can do better than proportional allocation. That is, we can achieve less sampling error by allocating the sample items between strata in an optimum fashion. Note the allocation of sample items in the railroad waybill example on page 321. The fifth stratum (revenue \$40 and over) contains $5\frac{1}{2}$ percent of the whole population of waybills and all (100 percent) of this stratum is included in the sample. On the other hand, the first stratum (revenue 0 to \$4.99) contains 56.0 percent of all waybills, but only 5 percent of this group is included in the sample.

Using optimum allocation we divide the total sample among the strata in such a way that we obtain the smallest sampling error for a given size of sample. The standard error is a function not only of the sample size within each stratum, but also of the variability of these items. To achieve optimum allocation we assign in proportion to both the size of the stratum and the standard deviation within the stratum.

The formula is thus

$$m_i = n \frac{M_i s_i}{\sum M_i s_i}$$

$$m_i = n \frac{M_i s_i}{\sum M_i s_i}$$

⁴ This is commonly called a *self-weighting sample*.

where n is the total sample size; M_i refers to the total number of items in the i th stratum; m_i is the sample size in that stratum; and s_i is an estimate of σ_i , the standard deviation of the items in the i th stratum.

To illustrate this, consider the example on page 322.

Table 14-4
STRATIFIED SAMPLE OF INCOMES—
OPTIMUM ALLOCATION

Stratum Number (i) (h_i)	Number of Items in Stratum (M_i) (w_i)	Standard Deviation of Items in Stratum (s_i) (σ)	Product ($M_i s_i$)
1	1,000	\$1,000	1,000,000
2	2,000	500	1,000,000
Total	3,000 = N		2,000,000

Table 14-4 shows the number of items (M_i) and the standard deviation (s_i), together with the product $M_i s_i$ and the total $\Sigma M_i s_i$.

Let us take a sample of $n = 600$ items as before. How should they be allocated to minimize sampling error? Using the above formula, the sample size for the first stratum should be

$$m_1 = (600) \frac{1,000,000}{2,000,000} = 300$$

and the sample size for the second stratum is also 300.

To review the formulas for sampling error with stratified sampling and to illustrate that optimum allocation does reduce sampling error, let us carry out the calculation of the standard error of the mean with optimum allocation.

When we use these sample sizes and other data from Table 14-4, the standard errors within each of the strata are

$$s_{\bar{Y}_i} = \frac{s_i}{\sqrt{m_i}} \sqrt{1 - \frac{m_i}{M_i}}$$

so that

$$s_{\bar{Y}_1} = \frac{1,000}{\sqrt{300}} \sqrt{1 - \frac{300}{1,000}} = \sqrt{2,333}$$

$$s_{\bar{Y}_2} = \frac{500}{\sqrt{300}} \sqrt{1 - \frac{300}{2,000}} = \sqrt{708.3}$$

And the standard error for the population mean is

$$s_{\bar{Y}_s} = \sqrt{\sum w_i^2 s_{Y_i}^2} = \sqrt{(1/3)^2(2,333) + (2/3)^2(708.3)} = \sqrt{574} = \$24$$

Note that this is quite a significant decrease over the previous allocation, which gave a sampling error of \$34.

Allocation of the Sample to Strata: Least-Cost Allocation. If there is a difference in cost of obtaining a sample item in the various strata, then we can introduce this cost into the considerations. The formula becomes

$$m_i = n \frac{M_i s_i / \sqrt{c_i}}{\sum (M_i s_i / \sqrt{c_i})}$$

where c_i is the cost of sampling one item in stratum i , and m_i , n , M_i , and s_i are as defined above.

To continue our example, suppose that it cost \$4 to obtain one item in stratum 1, and \$9 to obtain one item from stratum 2. Then for a sample of 600 items, we should allocate to stratum 1:

$$m_1 = (600) \frac{1,000,000/\sqrt{4}}{(1,000,000/\sqrt{4}) + (1,000,000/\sqrt{9})} = 360$$

Similarly, the allocation to the second stratum is $m_2 = 240$.

Thus, because it is cheaper to sample in stratum 1, a larger sample should be taken in that stratum than under the optimum allocation above.⁵

There are a few "loose ends" that need to be discussed before we can leave stratified sampling. The first is the question: How many strata and how should they be determined? Oftentimes, the number and the boundaries of strata are determined by administrative convenience. Certain geographic areas, such as counties or states, form natural boundaries. However, there are times when the survey designer can set the number of strata. Then, how many strata should he make? Let us first point out that as long as we can select strata that differ somewhat from each other (with different means or standard deviations for the varia-

⁵ We could go one step further and ask: Given C dollars, how many items should we take and how should they be allocated? The total sample size is

$$n = C \frac{\sum (M_i s_i / \sqrt{c_i})}{\sum (M_i s_i \sqrt{c_i})}$$

Then the allocation to strata can be done as above. See Edward C. Bryant, *Statistical Analysis*, 2d ed. (New York: McGraw-Hill, 1966).

ble measured) we can continually increase precision. That is, under this circumstance, the larger the number of strata the better. However, in any actual situation, we do not know the content of all possible strata, and some point is reached when we cannot be sure we are breaking the population into strata that differ from each other. At this point, the use of more strata does not increase the precision. And remember that the more strata, the more computations are needed to compile our final estimates.

Stratification and Nonresponse. One method of handling nonresponse in a survey is to consider the population as made up of two strata, one being those who respond (e.g., those who reply to a mail questionnaire), and a second stratum of those who do not respond. When a survey is taken, the respondents can be used as one subsample. Then a subsample of the nonrespondents is taken by other means (e.g., by follow-up interviews). This subsample of nonrespondents is then used to provide estimates for the nonresponse stratum.

As an example, suppose that 1,000 mail questionnaires are mailed out and 520 are returned. Thus there are 480 nonrespondents in the sample. Suppose that 1 out of 4 of these are selected at random (120 in total), and interviewers are sent to obtain the desired answers. The total sample size would then be $520 + 120 = 640$. However, the values obtained for the 120 nonrespondents would have to be multiplied by 4 to assure them of the correct weight.

For the error formulas and further discussion on this type of sampling, more advanced texts are recommended.⁶

Ratio Estimation

In many business and economic surveys, it is important to estimate not the mean of a population but a *ratio*. As noted elsewhere, the ratio (including the proportion, percentage, fraction, or index number) is a basic summary measure for comparing two attributes, just as the mean is a basic measure for summarizing variables.⁷ For example, an accountant may wish to sample a firm's accounts receivable to determine the ratio of balances in overdue accounts to the total balance of all accounts.

A ratio can also be used to estimate a population mean or total. For example, a ratio is often employed to approximate the total number of

⁶ See Leslie Kish, *Survey Sampling* (New York: John Wiley, 1965), pp. 132, 217, 304, and 532-62 and other readings listed at the end of this chapter.

⁷ Ratios are described in Chapter 4, the binomial distribution in Chapter 8, inferences involving proportions in Chapter 13, index numbers in Chapter 18, and quality control of attributes in Chapter 25.

wild animals in a certain area or the number of fish in a lake. A known number of animals or fish are tagged and released in the area to be surveyed. After allowing sufficient time for them to mix with the group, a number of the animals or fish are caught. The ratio of the number tagged to the total number caught then yields an estimate of the total number of animals or fish. For example, suppose 1,000 fish are tagged and put in a lake, and subsequently 200 fish caught, of which 20 are found to be tagged. That is, there is a ratio of 10 fish for every tagged one in the sample. Since the total number tagged is 1,000, the total number of fish is estimated at 10 times the number tagged, or $10 \times 1000 = 10,000$ fish.

As another example, the ratio of persons per water meter (say 3 to 1) is often used to make intercensus estimates of a city's population, since the number of water meters is usually easily obtainable. Similarly, the ratio of number of children in public schools to total population is used to estimate current population, since the count of schoolchildren is readily known.⁸

The use of ratio sampling to estimate a population mean or total depends upon the availability of certain auxiliary data that are related to the variable we are estimating. In the above examples, the number of water meters and the number of schoolchildren were the auxiliary data needed to estimate the total population. If such data are available, then ratio sampling can be quite efficient in reducing sampling error.

Let us consider an example in detail. A company wishes to estimate its total inventory value at the end of each month. This would require a fairly large sample, since the values of different inventory items are likely to have a large standard deviation—that is, they probably range from a few cents up to hundreds of dollars. We might be able to achieve some improvement by stratification. An easier approach, however, would be to use ratio sampling.

We can take a random or systematic sample of items from the inventory, and compare their total current value with their value in the last annual inventory, as in Table 14-5. Then we multiply the percent change in this sample value by the total annual inventory value, which was taken on a 100 percent basis, to estimate the total current inventory.

⁸ The perils in this process are obvious. Trends in the makeup of a city's population may change the ratio over time. Hence, inaccurate estimates will be made if the ratio is not re-estimated periodically. At least one large city received a severe shock at the time of the 1960 census, when population estimated as above was quite different from official census figures.

Table 14-5

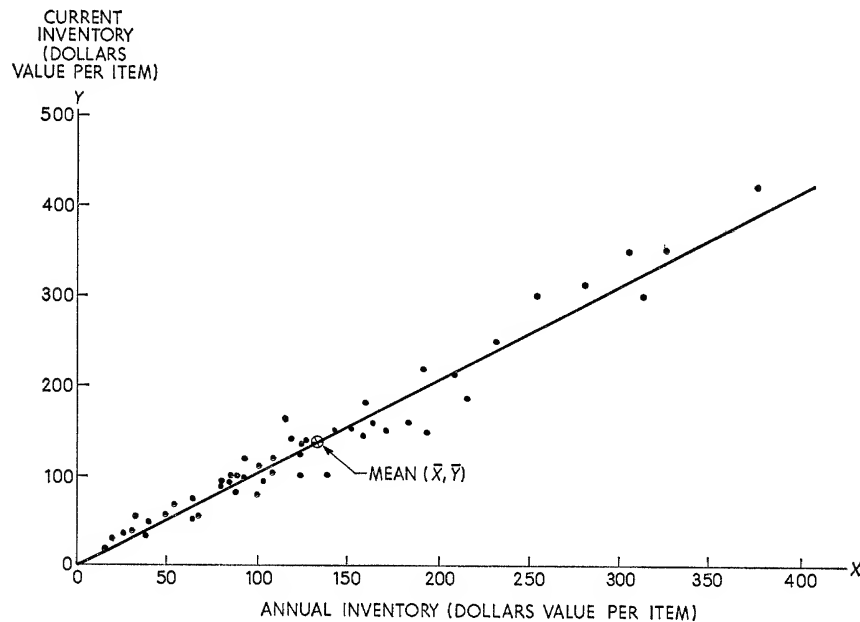
SAMPLE OF 50 ITEMS FROM THE INVENTORY RECORDS OF A COMPANY
VALUES FOR CURRENT AND ANNUAL INVENTORY (IN DOLLARS)

Item Number	Annual Inventory Value (X)	Current Inventory Value (Y)	Item Number	Annual Inventory Value (X)	Current Inventory Value (Y)
1	\$ 160	\$ 182	26	84	89
2	87	84	27	171	152
3	280	315	28	103	96
4	123	125	29	326	350
5	20	28	30	38	35
6	254	300	31	128	139
7	100	82	32	124	102
8	142	151	33	87	99
9	50	55	34	375	420
10	124	136	35	80	88
11	64	52	36	208	216
12	164	160	37	86	99
13	40	48	38	67	58
14	151	154	39	305	349
15	107	105	40	158	146
16	80	92	41	32	39
17	193	150	42	184	160
18	93	110	43	137	100
19	231	250	44	115	165
20	54	68	45	33	57
21	101	110	46	216	186
22	16	18	47	119	141
23	191	220	48	64	72
24	109	120	49	312	300
25	91	95	50	27	35
Totals				\$6,604	\$6,903

This ratio estimate of current inventory has a smaller sampling error than one based on a random sample of the current inventory alone, if the values of an item are related in the two periods. This relationship is shown in Chart 14-1. Here the dots showing the relation of annual to current inventory values by item cluster along a diagonal line. That is, a major item is likely to have a high value in both periods, while a minor item will have consistently low values. The sampling error of a ratio estimate depends on the standard deviation of the dots above and below this line, whereas the sampling error of the mean of a sample of current inventory items depends on the larger standard deviation of the Y values above and below their own mean. We shall carry out this illustration further after introducing notation and formulas.

Chart 14-1

RELATIONSHIP BETWEEN ANNUAL INVENTORY AND CURRENT INVENTORY
BY ITEMS, RANDOM SAMPLE OF 50 ITEMS



Notation and Formulas. Let Y denote the unknown variable that we are trying to estimate—the current inventory value. Let X denote the variable about which we have complete information—the dollar value per item at the last annual inventory. An inventory *item* here refers to a particular type of merchandise, such as a certain kind of spark plug or hammer. The *value* of an item is the number on hand times the cost per unit—not the cost of one unit alone. Thus, in Table 14-5, the value \$160 for item 1 might represent 80 hammers at a unit cost of \$2.

In our example, we take a sample of 50 items from the inventory and find their total value at each date; that is, ΣX (the annual inventory) and ΣY (the current inventory).⁹ Then we calculate the ratio R ,

$$R = \frac{\Sigma Y}{\Sigma X}$$

which is an estimate of the unknown true ratio relating the total populations of X and Y . In our example, the ratio compares current inventory

⁹ There is a slight problem that we have ignored in this simple example. Some items in either the annual inventory or the current inventory would be out of stock. The definition of the population would then have to be a list of all items in stock at both times.

with annual inventory. We can use this ratio to estimate the total of the Y values, as follows: $T_Y = RT_X$, where T_Y is the ratio estimate of the total of the Y population and T_X is the total of the X population, which is assumed to be known.

The *mean* of the Y values is estimated similarly: $\bar{Y}_R = R\mu_X$, where \bar{Y}_R is the ratio estimate of the true mean μ_Y of the Y population. This is to be distinguished from \bar{Y} , the mean of the sample items. The value μ_X is the mean of the X population, which is known.

Note that the sample mean \bar{X} generally will not be exactly the same as μ_X .

The total, of course, is N times the mean. That is, $T_X = N\mu_X$ and $T_Y = N\bar{Y}_R$, where N is the total number of items.

In our example (Table 14-5), the ratio of current inventory to annual inventory value for the sample of 50 items is

$$R = \frac{\Sigma Y}{\Sigma X} = \frac{6,903}{6,604} = 1.0453$$

That is, the inventory, by our estimate, increased 4.53 percent in value from the annual to the current inventory. Suppose the annual inventory totaled \$3,447,519. This is T_X . Then the total current inventory T_Y can be estimated as

$$T_Y = RT_X = (1.0453)(3,447,519) = \$3,604,000$$

Assume there were 24,167 inventory items in the annual inventory (i.e., $N = 24,167$), so that the mean value was

$$\mu_X = \frac{3,447,519}{24,167} = \$142.654 \text{ per item}$$

Then we could estimate the mean value per item of the current inventory as

$$\bar{Y}_R = R\mu_X = (1.0453)(142.654) = \$149.11$$

Note that this is different from \bar{Y} , the mean value of current inventory in the sample, which is $\$6903/50 = \138.06 . Thus, our estimate is considerably higher than we would have obtained from a simple random sample.

It may help to ponder this last statement. We are making a higher estimate using ratio sampling than we would have had we considered the sample as a simple random sample. Perhaps this is easiest seen if we consider the estimate of total current inventory. Our estimate from ratio

sampling is given above as \$3,604,000. The simple random sample estimate for a total is

$$T_Y = N\bar{Y} = (24,167)(138.06) = \$3,336,000$$

Thus, ratio estimation gives us an estimate that is \$268,000 above that obtained using a simple random sample estimate. Why is this so? The ratio estimate is higher precisely because we realize, from our knowledge of the X variable, that the sample has understated the population total. Note that \bar{X} (the value for the sample) is \$132.08, while the known population value is $\mu_X = \$142.65$. Hence, we adjust the value of \bar{Y}_R upward to correct this understatement. Of course, in some samples it will be necessary to adjust downward, for identical reasons.

It is also important to note that we are dependent upon a close relationship between X and Y for ratio sampling to be efficient. If this were lacking, there would be no sense in making the adjustment as we did above.¹⁰

Bias and the Ratio Estimate. Unfortunately, the ratio estimate is a biased estimator of the population ratio. That is, the average of the ratios obtained from many samples does not generally equal the true ratio in the population. However, this bias is quite small in large samples, and we can ignore it in this case.

The bias will be negligible even for small samples if the relationship between X and Y can be approximately described by a straight line through the origin. Examination of Chart 14-1 indicates that this is certainly true for our example of estimating current inventory from annual inventory.

The following general rule has been suggested by Cochran for determining when the bias in a ratio sample is negligible.¹¹

The bias in the ratio estimate, and the associated standard error, are of negligible size if

1. the sample size exceeds 30 and
2. both $\frac{s_Y}{\sqrt{n}\bar{Y}}$ and $\frac{s_X}{\sqrt{n}\bar{X}}$ are less than 0.10.

¹⁰ The ratio estimate is more efficient (i.e., has smaller sampling error for a given size sample) than simple random sampling if the X and Y variables are highly related. A measure of the relationship between X and Y is the correlation coefficient (see Chapter 22) defined as $r = \Sigma xy / \sqrt{\Sigma x^2} \sqrt{\Sigma y^2}$. Generally, the ratio estimate is more efficient than simple random sampling if $r > 1/2 \sigma_X \mu_Y / \sigma_Y \mu_X$.

¹¹ William G. Cochran, *Sampling Techniques*, 2d ed. (New York: John Wiley, 1963), p. 157.

Sampling Error of the Ratio Estimate. The amount of sampling error associated with the ratio R and the ratio estimates \bar{Y}_R , and T_Y can be estimated by the following formulas:

$$\text{Standard error of ratio} = s_R = \sqrt{\frac{\Sigma Y^2 + R^2 \Sigma X^2 - 2R \Sigma XY}{n(n-1)\bar{X}^2}} \sqrt{1 - \frac{n}{N}}$$

ΣXY is the cross-product term and is obtained by multiplying and then summing corresponding values of X and Y . The last term is the finite population correction and may be omitted if the sample is less than 5 percent of the population.

$$\begin{aligned} \text{Standard error of mean} &= s_{\bar{Y}_R} = s_R \bar{X} \\ &= \sqrt{\frac{\Sigma Y^2 + R^2 \Sigma X^2 - 2R \Sigma XY}{n(n-1)}} \sqrt{1 - \frac{n}{N}} \end{aligned}$$

$$\text{Standard error of total} = s_{T_Y} = N s_{\bar{Y}_R}$$

When the true mean μ_X is known, it should be used in place of \bar{X} in the above formulas.

To illustrate, let us continue the example of estimating total current inventory. The standard error for this estimate is, as above,

$$s_{T_Y} = N s_{\bar{Y}_R} = N \sqrt{\frac{\Sigma Y^2 + R^2 \Sigma X^2 - 2R \Sigma XY}{n(n-1)}} \sqrt{1 - \frac{n}{N}}$$

From Table 14-5, we can calculate the following:

$$\begin{aligned} \Sigma Y^2 &= 1,365,701 \\ \Sigma X^2 &= 1,227,238 \\ \Sigma XY &= 1,285,673 \end{aligned}$$

Recall also that

$$\begin{aligned} n &= 50 \\ N &= 24,167 \\ R &= 1.0453 \end{aligned}$$

Since the sample is a very small part of the total population, the finite population correction in the above formula can be ignored. Then:

$$\begin{aligned} s_{T_Y} &= (24,167) \sqrt{\frac{1,365,701 + (1.0453)^2(1,227,238) - 2(1.0453)(1,285,673)}{50(49)}} \\ &= (24,167) \sqrt{\frac{18,820}{2,450}} \\ &= 66,980 \end{aligned}$$

Thus, our estimate of total current inventory is \$3,604,000 with a standard error of \$67,000. This standard error is only 2 percent of the total, with a sample of 50 items, so ratio estimation is quite efficient in this case. For comparison, the sampling error obtained from simple random sampling is about \$314,000.¹²

Before using the standard error to determine confidence limits, we should check the rules on page 332 for determining if bias is negligible. Note that

1. sample size is greater than 30 ($n = 50$),
2. $\frac{s_Y}{\sqrt{n}\bar{Y}} = \frac{91.76}{\sqrt{50} \cdot 138.06} = 0.094$ which is less than 0.10
and
 $\frac{s_X}{\sqrt{n}\bar{X}} = \frac{85.11}{\sqrt{50} \cdot 132.08} = 0.091$ which is also less than 0.10.

Hence, we will not worry about bias in the estimates of T_Y and s_{T_Y} .

Cluster Sampling

Cluster sampling is the procedure by which a population is divided into several groups or clusters. A number of these clusters are then drawn into the sample and a subsample (possibly 100 percent) of elements is selected from each of the specified clusters. Thus, we are sampling at two stages: the first stage where a sample of clusters, called *primary sampling units*, is drawn and a second stage in which individual elements, called *elementary sampling units*, are taken from the selected clusters.

We shall discuss only two-stage sampling, but there is no reason why three or even more stages could not be employed. For example, in sampling a city we could define the primary unit as the block, the secondary unit as the dwelling unit, and the tertiary unit as the individual. When each cluster is contained in a separate geographic area, cluster sampling is also called *area sampling*. The main advantage of cluster sampling is that it reduces the cost per elementary sampling unit. To understand this, suppose we were taking a sample of business establishments in a certain county. If a simple random sample were

¹²To see this:

$$s^2_Y = \frac{\Sigma Y^2 - \bar{Y}\Sigma Y}{n - 1} = \frac{1,365,701 - (138.06)(6,903)}{49} = 8,421.9$$

$$s_Y = 91.76$$

$$\text{Estimate of standard error of mean} = s_{\bar{Y}} = \frac{s_Y}{\sqrt{n}} = \frac{91.76}{\sqrt{50}} = 12.977$$

$$\text{Estimate of error in total} = s_{T_Y} = N s_{\bar{Y}} = (24,167)(12.977) = 313,600$$

selected, the establishments in the sample would be scattered widely over the whole county. It would take interviewers a considerable amount of travel time to obtain the desired results. On the other hand, suppose the county were first broken up into geographic areas (clusters), and a sample of the clusters was taken. Then a subsample of the establishments within the selected areas is determined. With this procedure considerable travel time for the interviewer would be saved since all of the establishments sampled will be clustered in the areas selected rather than spread randomly over the county.

A second advantage of cluster sampling is that it can be used sometimes where other methods are not applicable. For example, in selecting the sample of business establishments above, a complete list of all the establishments may not be available or feasible. However, it would be relatively simple to divide the county into geographic areas and select a number of these clusters as a sample. Business establishments could be listed and sampled within the selected areas without great difficulty. That is, we would have to prepare lists only within the selected areas.

On the other hand, cluster sampling is relatively inefficient. The results of a cluster sample are usually *not* as precise as those of a random sample of *the same size*. They can be made equally or more precise only by taking a larger sample. The cost of conducting a survey, however, may still be lower. For example, instead of spending \$10,000 to interview a random sample of 1,000 householders at an average cost of \$10 each, one might get better results for \$9,000 with a cluster sample of 1,500 householders costing only \$6 each.

Serpentine Numbering and Systematic Selection. A recommended method of selecting the clusters in area sampling is to number the primary sampling units in a *serpentine* sequence, following a winding path similar to that of a snake (see diagram). For example, in a study of household incomes, the numbering of city blocks should follow a sequence of blocks having about the same average household income. All blocks in such an area should be numbered before proceeding to a lower-income or higher-income area. After the block map has been numbered, the desired number of blocks should be chosen by *systematic* selection (e.g., every tenth block) with a random start, as explained previously.

SERPENTINE NUMBERING
OF CITY BLOCKS

1	2	3	4	5
10	9	8	7	6
11	12	13	14	15

Table 14-6

RESULTS OF A SAMPLE TO ESTIMATE AVERAGE HOUSEHOLD INCOME IN A CERTAIN CITY

	Block No. (Determined by Random Number)	Number of Households in Block	Average Income of 3 households in Block (\$000)	Estimate of Total Income of all Households in Block (\$000)
i		M_i	\bar{Y}_i	$T_i = M_i \bar{Y}_i$
1	643	45	10.7	480.0
2	346	63	5.7	357.0
3	960	52	7.3	381.3
4	236	54	11.7	630.0
5	730	54	9.6	522.0
6	376	65	5.3	346.7
7	25	71	6.7	473.3
8	203	62	6.3	392.7
9	639	66	5.0	330.0
10	91	55	7.7	421.7
11	505	61	11.7	711.7
12	922	71	9.0	639.0
13	310	57	6.0	342.0
14	459	73	7.7	559.7
15	595	67	11.0	737.0
16	936	67	9.7	647.7
17	879	63	8.3	525.0
18	707	53	8.3	441.7
19	733	66	9.3	616.0
20	166	49	11.7	571.7
21	750	65	7.0	455.0
22	550	59	6.3	373.7
23	425	60	9.7	580.0
24	576	54	10.3	558.0
25	360	57	11.7	665.0
26	721	49	8.3	408.3
27	685	55	10.7	586.7
28	440	56	8.3	466.7
29	297	47	6.3	297.7
30	107	71	7.3	520.7
Total		1,787		15,038.0

This area sampling design achieves all of the advantages of geographic stratification when blocks in one stratum are numbered before proceeding to another stratum. However, stratification by some other characteristic, such as block size, is sometimes advisable.

Subsampling. After the primary sampling units have been chosen, elementary sampling units are selected from each of these clusters. The

selection may be a complete census of the cluster (e.g., all the houses in the block) or a random or systematic sample (e.g., every fifth house).

The cost per interview for a subsample is higher than that for a complete census of the selected clusters. The choice between these alternatives depends in part on the complexity of the interview and the availability of lists. If the questionnaire is simple and no list of elementary sampling units (e.g., households) is available, it is usually cheaper to take a complete census of the selected clusters (e.g., blocks); when a lengthy interview is required, the advantages of subsampling justify the cost of listing and sampling the elementary sampling units.

Let us consider a single example to illustrate the concepts involved in cluster sampling. Suppose we were interested in estimating the average family income in a certain city. There are 997 blocks in the city, and they are numbered in the serpentine fashion described above. Thirty blocks are selected at random. In each selected block, the number of households is determined and a sample of three households is selected. An interviewer is sent to the head of the selected households and total household income determined. The results are shown in Table 14-6.

In this example, the primary sampling unit is the city block and the secondary unit is the household. Note that the number of households in the whole city may not be known. We need to know only the number of households in each of the blocks selected, and this information may be readily obtainable.

Notation and Formulas. Before we can convert the data contained in Table 14-6 into an estimate of the average income in the city, it will be necessary to present the formulas and symbols used in them. Let:

M = the total number of secondary units (households) in the whole population

N = the number of primary units (blocks in this example) in the population

n = the number of primary units (blocks) in the sample

M_i = the number of secondary units in the i th primary unit—the number of households in the i th block

\bar{Y}_i = the average of the sampled secondary units in the i th primary unit—average income in the i th block

m_i = the number of secondary units sampled in the i th primary unit—number of households sampled in the i th block

$T_i = M_i \bar{Y}_i$ be the estimate of the total for the i th cluster—total income in the i th block

A simple estimate of the mean of the population (average income per household) is

$$\bar{Y}_c = \frac{\sum M_i \bar{Y}_i}{\sum M_i} = \frac{\sum T_i}{\sum M_i}$$

Note that this formula does not involve M , the total number of all secondary units (households). Only the M_i , the number of households in the sampled blocks, is required.

The cluster sample estimate \bar{Y}_c is biased, but the bias is small if a fairly large number of primary units (blocks) are sampled.¹³

An estimate of the sampling error of the cluster estimate \bar{Y}_c is

$$s_{\bar{Y}_c} = \sqrt{\left(\frac{N}{M}\right)^2 \frac{\sum (T_i - M_i \bar{Y}_c)^2 (1 - n/N)}{n(n-1)} + \frac{(N/n) \sum M_i^2 s_{\bar{Y}_i}^2}{M^2}}$$

where $s_{\bar{Y}_i}$ is the standard error of the estimate of \bar{Y}_i in the i th cluster (the error associated with the estimated average income in a block),

$$s_{\bar{Y}_i} = \frac{s_i}{\sqrt{m_i}} \sqrt{1 - \frac{m_i}{M_i}}$$

where s_i is the standard deviation of the items sampled in the i th cluster.

When M is not known, use the estimate $N \sum M_i / n$ instead.

Note that the equation for $s_{\bar{Y}_c}$, the standard error of the cluster estimate, has two parts. The first term is roughly related to the variability *between* cluster totals, and the second term, to the variability *within* clusters. The first term generally is the larger. *In fact, if the sampled clusters represent a small fraction of the total number (n/N less than about 0.05), the second term becomes small and can be ignored in calculations.*

In our example (Table 14-6) of sampling incomes in a city, the estimate of the mean income per household is

$$\bar{Y}_c = \frac{\sum T_i}{\sum M_i} = \frac{15,038.0}{1,787} = 8.415 \text{ thousands of dollars}$$

and the estimated sampling error of this mean is

$$s_{\bar{Y}_c} = \sqrt{\left(\frac{N}{M}\right)^2 \frac{\sum (T_i - M_i \bar{Y}_c)^2}{n(n-1)}}$$

using only the first term and ignoring the finite population correction $(1 - n/N)$ since n is only 3 percent of N . Here $N = 997$; $n = 30$; and M is estimated as

¹³ An unbiased estimate is also available if M is known. However, the unbiased estimate is generally less efficient than the biased estimate above. See Cochran, *Sampling Techniques*, pp. 300-305, for more details.

$$\frac{N}{n} \Sigma M_i = \frac{997}{30} (1,787) = 59,388$$

and

$$\Sigma(T_i - M_i \bar{Y}_c)^2 = 437,811 \text{ (calculation not shown)}$$

$$s_{\bar{Y}_c} = \sqrt{\left(\frac{997}{59,388}\right)^2 \left(\frac{437,811}{30(29)}\right)} = \sqrt{0.1418}$$

$$= 0.377 \text{ thousands of dollars}$$

This is a fairly large sampling error—about 4.5 percent of the mean—considering the size of the total sample (90 households). A simple random sample of 90 households would have been more accurate. However, the 90 households in the cluster sample would be considerably cheaper to survey than the equivalent simple random sample. Furthermore, taking a simple random sample would have been impossible to do without first compiling a complete list of all households in the city—quite a task!

The method described above is one way in which cluster sampling can be formulated. Other methods are useful for different situations. For example, when the primary units or clusters vary greatly in size, a technique may be used which will make the probability of selecting a cluster proportioned to the size of the cluster. In addition, three or even more stages may be used, as noted earlier. These require more complicated formulas, but the basic ideas illustrated above are the same. Note that cluster sampling is used in conjunction with other sample types, such as random or systematic samples, which are needed to select both the primary and secondary sampling units.

We have skirted over some of the major problems associated with cluster sampling, such as: How many clusters? How large should they be? How many units should be in the subsample from the cluster? How do we compare the cost of a cluster sample with other methods? These questions have been left to advanced texts.

Replicated Sampling

Replicated sampling is a technique of selecting independent subsamples of the population (sometimes called “interpenetrating” subsamples). For example, instead of a random sample of 200 elements from some population, one might divide the 200 into 10 subsamples, each consisting of 20 elements. Or the 90 households sampled in Table 14-6 might be broken into 3 subsamples, each consisting of 30 items—1

household from each of the 30 clusters. The subsamples are structured exactly the same—that is, they are *replicas* of each other. With replicated sampling, the overall estimate of the mean is the mean of the individual subsample estimates.

One main use of replicated sampling is in determining the sampling error for complicated sample designs.¹⁴ Consider, for instance, the sampling error formula for cluster sampling on page 338. If we added a third stage in the cluster design plus stratification, the formula could become quite unwieldy. With replicated samples, this sort of calculation is not necessary. Also, for systematic sampling, the sampling error is difficult to estimate unless the elements in the population are in random order. Again, replicated samples may be used to make a simple estimate of the sampling error.

Suppose that k replicated samples are drawn and for each a mean \bar{Y}_j is calculated. Each \bar{Y}_j is an estimate of the population mean. The overall replicate sampling estimate of the mean is

$$\bar{Y} = \frac{\sum \bar{Y}_j}{k}$$

and the estimated sampling error is

$$s_{\bar{Y}} = \sqrt{\frac{\sum (\bar{Y}_j - \bar{Y})^2}{k(k-1)}}$$

In words, the standard error $s_{\bar{Y}}$ is determined only from the variance of the sample means \bar{Y}_j themselves,¹⁵ thus avoiding all calculations of variances within and between clusters, within strata, etc.

The number of replications k to make depends upon various factors in the design. Deming suggests $k = 10$ as a good number for a wide variety of applications.¹⁶

NONPROBABILITY SAMPLING

Nonprobability sampling includes any method of sampling which does not satisfy all requirements of a probability sampling design. This

¹⁴ Replicated sampling is also used to estimate possible measurement error in the survey. Thus, if each subsample is taken from the reports of a separate interviewer, a replicated sample could reveal interviewer bias. The use of replication in nonprobability sampling is described below.

¹⁵ The estimated sampling error $s_{\bar{Y}}$ has $k - 1$ degrees of freedom. In determining confidence intervals, therefore, it may be necessary to use the t distribution.

¹⁶ W. Edwards Deming, *Sample Design in Business Research* (New York: John Wiley, 1960), Chap. 21. Chapters 6–15 present a thorough treatment of replicated sampling designs.

may involve selection of a sample according to personal convenience (to minimize cost) or expert judgment (to increase precision in certain small samples) or under conditions where no complete list is available for objective selection (e.g., a survey of executives who influence corporate buying policy on industrial equipment). Nonprobability sampling methods are important in business and economic research despite the disadvantage that the precision of their results cannot usually be measured objectively. Two principal types of nonprobability sampling are *quota sampling* and *judgment sampling*.

Quota Sampling

A *quota sample* is one in which the interviewer is instructed to collect information from an assigned number, or quota, of individuals in each of several groups—the groups being specified as to age, sex, income, or other characteristic—much like the strata in stratified sampling. Subject to these controls, however, the individuals selected in each group are left to the interviewer's choice rather than being determined by probability methods.

For example, the McGraw-Hill Publishing Company carries out numerous attitude surveys among executives who read industrial magazines to aid the McGraw-Hill management in the conduct of its own publications. Readers are asked what journals and other sources of information they use, what topics interest them most, and similar questions. Interviews are conducted by "resident investigators"—mostly women living in the survey area. In one such survey, covering chemical process industries, the company's Research Department had a complete list of plants but no comprehensive list of individual executives. A stratified, systematic sample of plants was first selected in each area. Given this list, each investigator was instructed to locate and interview a specified number of engineers, production men, etc. who had some influence on the company's purchasing policy. The investigator would typically interview one to three engineers etc. in each plant and continue to other plants in the area until her quota was completed. This quota method was considered by the director of marketing research to be the only feasible way of conducting an industrial survey when the population of respondents could not be identified.

Quota sampling is popular in market surveys and public opinion polls because it is cheaper per elementary sample unit than random sampling and, when carefully controlled, has many of the advantages of stratified random sampling. However, it is subject to two important sources of error: (1) the quotas set for the interviewer represent a

rather crude stratification plan for the population, being based on only a few broad criteria, such as age (young, middle-aged, or old) and income (low, middle, or high); (2) since the interviewer is free to select individuals within a quota, he may choose people in convenient locations who may not be typical of the class of the population they have been chosen to represent. For example, in a survey of the number of young children by households, the method of interviewing women who happen to be at home would be apt to yield a sample with too large a proportion of women with young children, because such women are more likely to be at home during the hours in which the interviewing is done than are other women. Therefore, interviewers must be carefully trained to avoid such pitfalls.¹⁷

Quota sampling has been popular in preelection polls since 1936 when such samples yielded results far superior to those of a much larger mail questionnaire conducted by the *Literary Digest*. The larger poll was in error mainly because the sample was taken from telephone books and automobile registration lists, so that it contained too many voters from higher-income groups. This bias persisted despite the very large size of the sample.

Most presidential preelection polls have been successful since then, except in 1948, when they predicted a victory for Dewey rather than Truman. It is not certain to what extent this error is attributable to the quota sampling methods used or to other factors, such as improper interviewing techniques, the difference between what voters say and how they vote, or the shift of voters toward Truman after the preelection polls closed. In 1960, most polls correctly predicted the close Kennedy victory. The Gallup Poll, for example, forecast his margin at 52 percent of the combined Kennedy-Nixon vote, as compared with the actual figure of 50.1 percent. In 1964, Johnson's one-sided victory over Goldwater was correctly predicted, but in 1966 the polls underestimated the Republican comeback in many state and congressional elections.

It is often argued that all large-scale surveys should be based on a probability sampling design because of its greater objectivity. On the other hand, since a much larger quota sample can be taken for the same

¹⁷ Sometimes the sample is chosen so that the *average* age, income, or other pertinent characteristic of the individuals selected is equal to the average for the population. This is sometimes called "controlled" or "purposive" sampling. However, this control does not necessarily mean that the sample will be typical in other respects, such as in buying habits. Furthermore, this method is more difficult to administer than the simpler quota method, so it is used less frequently.

cost as a smaller probability sample, and because population lists may be unavailable, quota samples are still favored in some circumstances.

Judgment Sampling

A judgment sample is one which is selected according to someone's personal judgment. A judgment sample may be superior to a probability sample (1) in very small-scale surveys, (2) in "pilot studies" which precede major surveys, or (3) in constructing index numbers. Also, they are often less costly than probability samples. Unfortunately, however, judgment samples may be biased, and it is difficult to assess the validity of their results.

Examples of judgment samples in small-scale surveys include the choice of a single plant (i.e., a sample of one) in which to try out a new personnel policy or the choice of a few typical cities in which to make a market survey. A recent survey of consumer preferences for shampoo was conducted in San Jose, California, since this city was considered to be typical of the western market for this product. Such a judgment selection was probably superior to choosing a single city at random from a list of all cities in the West. This advantage of judgment selection, however, rapidly diminishes as the size of the sample increases because there is a steady increase in the precision of a probability sample, while the bias of the investigator persists in judgment sampling.

In pilot studies, which are designed to pretest a questionnaire to be used in any large survey, emphasis is placed on detecting unforeseen difficulties, which can be overcome by revising questions, rearranging the schedule, or training interviewers. For this purpose, respondents in a pilot study are often chosen on a judgment basis in such a way as to *overrepresent* types of individuals most likely to cause difficulties.

Another type of statistical work in which judgment selection is usually preferred to probability selection is that of index number construction (described in Chapter 18). Consider the problem of choosing a sample of 400 goods and services that make up the *Consumer Price Index* of the U.S. Bureau of Labor Statistics. There should be sample items for each of several broad classes of expenditures made by the typical family. These items should be representative of their classes with respect to price movements, and they should have some importance in themselves. In view of these and similar difficulties, items used in the construction of index numbers are usually chosen according to the judgment of experts in the field. Probability selection in such cases is applied only to classes in which there are a great many items of the same order of importance.

Accordingly, judgment selection is recommended for samples which are too small for the advantages of more objective methods, for pilot studies in which certain types of bias may actually be desirable, and for the selection of components in index numbers. Objective methods of selection, however, are necessary to attain a high degree of reliability in most large samples.

Standard Errors of Nonprobability Samples

The precision and standard errors of probability samples can be measured because the sample statistics follow the laws of chance (e.g., the means of large random samples follow the normal distribution) so that we can set confidence limits or test hypotheses with known probabilities. The standard error of a nonprobability sample, on the other hand, has no such significance, since sampling variation reflects unknown errors of judgment rather than chance.

However, if we take a *replicated* sample from the items in a nonprobability sample, all of the subsamples reflect about the same judgment factors, since they are replicas in their design. The subsample means, therefore, will vary because of numerous *chance* factors, and hence may follow a normal distribution. Hence, the standard error of the replicated sample is claimed to have some probability significance.

As an example, the standard error has been computed for a replicated sample of the items priced in the *Consumer Price Index*,¹⁸ using pairs of subsamples for different items (e.g., different models of cars priced) and different stores and different cities to provide a total of 732 city-group relatives. Each of these subsamples is carried forward monthly from a base in December 1963. Then, since many independent factors affect the dispersion of the 732 means, they are believed to be normally distributed, and the standard errors are computed for each month by the formula given above for replicated samples. Thus, the index for transportation cost in October 1964 was 100.48 (December 1963 = 100), with a standard error of 0.19 points. The validity of this figure is controversial. Nevertheless, replicated sampling provides a possible means of making a rough estimate of the precision of nonprobability samples in general.

¹⁸ See M. Wilkerson in *1964 Proceedings of the Business and Economic Statistics Section* (American Statistical Association), pp. 220-33; also J. C. Sawhill in *1963 Proceedings*, pp. 9-20; and P. J. McCarthy in *1961 Proceedings*, pp. 264-70; as well as P. J. McCarthy in *The Price Statistics of the Federal Government* (National Bureau of Economic Research, 1961), pp. 197-232.

SUMMARY

Information obtained from samples is indispensable in modern business and economic research. It is important, therefore, to plan sample surveys in such a way as to obtain the desired information with maximum precision and minimum cost of time and effort.

Probability sampling includes all methods (such as simple random sampling, stratified random sampling, systematic selection, and cluster sampling) in which there is a known probability of being selected for each individual in the population. Nonprobability sampling includes all other methods, such as quota and judgment sampling. Probability sampling methods have a basic advantage in that the precision of their results can be measured objectively and compared as between different sample designs. This is especially important in very large samples.

A *simple random sample* of n units is one selected from the population in such a way that each combination of n units has an *equal* probability of being selected. A table of random numbers is usually used to select items at random.

Systematic sampling is the process of taking observations at equal intervals in a list. When nearby parts of a population are alike, systematic sampling with a random start is superior to simple random sampling in spacing the sampling units more evenly over the population.

A *stratified random sample* is one in which the population is divided into fairly uniform groups or strata. Then a random sample is drawn from each selected stratum. If the various strata can be made more homogeneous than the population as a whole, a stratified sample will yield more precise results than a simple random sample of the same size.

The total sample must be apportioned to the various strata. *Proportional allocation* assigns the sample elements to the strata in the same proportions of the total sample as they occur in the population. If the strata differ considerably in variability, then *optimal allocation* will improve the estimate and should be used. Optimal allocation assigns the sample to the strata in proportion to the strata size and standard deviation within the strata. If the cost of sampling varies considerably from stratum to stratum, then *least cost allocation* should be employed to maximize precision relative to cost.

Stratification of a population into respondents and others, and subsampling from the nonrespondents, is one method for dealing with nonresponse in surveys.

Ratio estimation focuses on proportions rather than on means. A ratio estimate may also be used to estimate the mean (or total) of one

population, using the ratio between the variable to be estimated and an auxiliary variable that is related to the first and about which complete information is available.

The efficiency of the ratio estimate depends upon the relationship between the two variables used in the estimate. If the two variables are strongly related, the ratio estimate can have a much smaller sampling error than a simple random sample. The ratio estimate is biased (the average of many ratio estimates would not give exactly the population value), but the bias is negligible if the sample size is large.

Cluster sampling involves (1) selecting groups or clusters as primary sampling units and (2) taking a census or sample of "elementary sampling units" or secondary units within these groups. Cluster sampling is called *area sampling* when the cluster falls in some geographic division, such as a city block. A cluster sample yields less precise results than a simple random sample of the same size, but the cost may be much less. The clusters are often chosen by systematic selection from a map on which areas are numbered by *serpentine* order.

There are several methods of cluster sampling. One is to sample the primary units with equal probability and subsample secondary units. Formulas and an illustration of this technique are presented. If the primary units vary greatly in size, they may be selected with probability proportionate to size. Other methods are also available.

The technique of *replicated sampling* involves drawing several independent subsamples from the population, all using the same sample design. The use of replicated samples makes the estimation of sampling error relatively easy.

Nonprobability sampling (including quota sampling and judgment selection) is the selection of a sample according to personal choice, expert judgment, or under conditions where lack of data prevents a probability selection. It is sometimes recommended when probability sampling is not feasible.

In *quota sampling* the investigator may choose the respondents from a quota or assigned number of individuals in each designated class. A quota sample is cheaper per unit than stratified random sampling and is popular in market surveys and public opinion polls, despite the serious pitfalls inherent in this method.

Judgment sampling is the selection of a sample based on expert judgment. It is recommended for surveys in which the sample is very small, for pilot studies preceding larger surveys, and for most economic index numbers.

The standard error of a nonprobability sample may possibly be estimated by replication, as in the case of the Consumer Price Index.

SUMMARY OF FORMULAS

Type of Sampling	Estimate of Population Mean	Estimate of Standard Error
Simple Random Sample	$\bar{X} = \frac{\sum X}{n}$	$s_{\bar{X}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$
Stratified Sampling	$\bar{Y}_s = \sum w_i \bar{Y}_i$	$s_{\bar{Y}_s} = \sqrt{\sum w_i^2 s_{\bar{Y}_i}^2}$
Ratio Estimation	$\bar{Y}_R = R \mu_x = \frac{\sum Y}{\sum X} \mu_x$	$s_{\bar{Y}_R} = \sqrt{\frac{\sum Y^2 + R^2 \sum X^2 - 2R \sum XY}{n(n-1)}} \sqrt{1 - \frac{n}{N}}$
Cluster Sampling	$\bar{Y}_c = \frac{\sum M_i \bar{Y}_i}{\sum M_i}$	$s_{\bar{Y}_c} = \sqrt{\left(\frac{N}{M}\right)^2 \frac{(\sum (T_i - M_i \bar{Y}_c)^2)(1 - n/N)}{n(n-1)}} + \frac{(N/n) \sum M_i s_{Y_i}^2}{M^2}$
Replicated Sampling	$\bar{Y} = \frac{\sum \bar{Y}_j}{k}$	$s_{\bar{Y}} = \sqrt{\frac{\sum (\bar{Y}_j - \bar{Y})^2}{k(k-1)}}$
Optimal Allocation to Strata	$m_i = n \left(\frac{M_i s_i}{\sum M_i s_i} \right)$	Size of Subsample in Stratified Sampling
Least Cost Allocation to Strata	$m_i \neq n \frac{\frac{M_i s_i}{\sqrt{c_i}}}{\sum \left[\frac{M_i s_i}{\sqrt{c_i}} \right]}$	

PROBLEMS

1. Comment on the following statements:
 - a) Sampling errors are due to improper methods of selecting a sample.
 - b) Survey results may be made as accurate as necessary by increasing the size of the sample.
 - c) A complete census is always preferable to a sample, if time and money permit.
 - d) Probability sampling should be used in all large-scale surveys, to obtain valid results.
 2. Distinguish between:
 - a) Probability sampling and nonprobability sampling.
 - b) Probability sampling and simple random sampling.
 - c) Stratified sampling and quota sampling.
 - d) Proportional and nonproportional sampling in stratified samples.
 - e) Primary and elementary sampling units in cluster sampling.
 3. You wish to conduct a survey of students in a university to determine which facilities they prefer (e.g., swimming pool, bowling alley, cafeteria) in a new student union building that is being planned. Compare the advantages of each of the three pairs of sampling methods in Problems 2 (a), 2 (c), and 2 (d) above for this purpose.
 4. Time Inc. made a survey of college graduates to determine their success and satisfaction in life, as related to their education record, and various other characteristics that would aid *Time Magazine* in analyzing its readership. Using lists supplied by colleges, *Time Magazine* sent questionnaires to all 15,700 graduates whose names began with "Fa" (Farley, Farmer, etc.). Over 9,500 replies were received.
 - a) What method of sample selection is this?
 - b) What sources of error might distort the results?
 - c) Suggest another method of selecting a sample of this size, that seems preferable to you, and show why this method should reduce the errors of response without greatly increasing the cost of the survey.
 5. Each student is to select a sample of 25 values of a quantitative variable and compute the average by adding the values and dividing the sum by 25. To insure comparability of results obtained by the various members, the class should agree on the choice of variable and the method of selection to be used. Problems to be considered include:
 - a) Are the data readily available?
 - b) If values are recorded on cards, might the cards be shuffled to arrange them in random order?
 - c) Are the values listed and numbered in order so as to facilitate selection by means of a table of random numbers?
 - d) Would systematic selection be effective?
-

- e) What strata might be constructed for stratified sampling?
6. As a distributor of major household appliances, you wish to survey the potential market for new appliances in your town by interviewing a sample of householders. Plan a cluster sample of the area as follows:
- Secure an up-to-date map of the town or one district of a larger city.
 - Number the blocks, or equivalent area, in serpentine fashion so as to follow a sequence of blocks having about the same household incomes.
 - Choose a systematic sample, with random start, of 20 blocks on this map.
 - Visit the tenth block selected (as an example) and list all house or apartment numbers around the block.
 - Select a random sample of six houses or apartments from this block, using a table of random numbers.
 - Comment briefly on the validity of this procedure for the problem at hand.
7. A population is divided into two strata, and a sample is taken from each stratum as follows:

	Stratum 1	Stratum 2
Number of elements in stratum = M_i	1,000	4,000
Number in sample = m_i	100	225
Stratum sample mean \bar{Y}_i	85	75
$\sum y_i^2$ in stratum, where $y_i = (Y_i - \bar{Y})$	9,900	89,600

- Estimate the mean for the whole population.
 - Estimate the standard error of the mean of the whole population.
8. An election is being held in a certain plant to determine if the workers should be represented by a union. To estimate beforehand the preference of the workers, management hires a consulting firm to take a sample of workers. The results are shown below in the table.

Department	No. of Workers in Department	No. of Workers in Sample	No. of Workers in Sample Voting for Unionization
1	5,000	100	60
2	5,000	50	20
Totals	10,000	150	80

- What estimate should management make of the proportion of workers in the whole plant voting for unionization?
 - What is the sampling error of this estimate?
- Hint:* The standard error of the proportion in each stratum is

$$s_{\bar{p}_i} = \sqrt{\frac{p_i q_i}{m_i}}$$

Use this in the same fashion as the standard error $s_{\bar{Y}_i}$.

9. As a dealer in retail hardware you are considering buying out the inventory of a merchant who is going out of business. You have a list of the items that he carried in stock but no exact inventory count has been made. There is the added problem of evaluating the worth of these items since many are obsolete or so old and damaged that they are worthless. Accordingly, you decide to take a sample of the items, check the count, and value carefully the sampled items.

The inventory is broken down into three product groups, including a special group for high-valued items. The number of items in each group is shown below. In addition, you make the following rough estimates of the standard deviations of the values of the items for each product group.

Product Category	Number Items in Product Category	Approximate Standard Deviation
High-value items.....	100	\$120
Paints and paint products.....	400	20
General hardware.....	500	10
Total	1,000	

Suppose you were considering a total sample of 50 items.

- How would you allocate the items by proportional allocation? By optimum allocation?
 - Estimate the standard error of the sample mean using proportional allocation and using optimal allocation.
10. A market research firm conducted a survey to estimate the percent of the population in a certain city that preferred a particular brand of soft drink. In order to obtain additional information, the city was divided into three areas, corresponding roughly to the high-, medium-, and low-income groups, respectively. A sample was then taken in each area. The results are shown in the table.

Income Area	Approximate Number of Consumers	Number Sampled	Number Preferring Brand X	Percent Preferring Brand X
High.....	20,000	80	16	20
Medium.....	120,000	150	75	50
Low.....	60,000	120	72	60
Totals.....	200,000	350	163	

- Make an estimate of the overall percent of consumers who prefer Brand X.
- How much sampling error is associated with the above estimate? Compute a 95 percent confidence interval about your estimate above.

Note: Recall that the formula for the sampling error of a proportion is

$$s_{p_s} = \sqrt{\frac{pq_s}{n}}$$

This is equivalent to the $s_{\bar{y}_i}$ in the formula for the estimate of the standard error in stratified samples.

- c) If you were to design a survey to be taken for a similar product (i.e., the percents within the various groups are expected to be the same as above), how would you allocate a proposed sample of 400 among the three income groups? (Let $s_i = \sqrt{p_i q_i}$)

11. The A & B Sporting Goods Company was interested in estimating the annual expenditures for camping gear for the 100,000 family units in the San Jose, California, area. In order to obtain information for the designing of a sampling plan, a pilot sample of 100 family units was chosen at random. The estimated annual expenditures for camping gear (U_i) and the annual family income (Z_i) were obtained for each family unit. A summary of these numbers is shown below:

$$\begin{aligned}\bar{U} &= \text{average expenditure} = \$26 \\ \sum U_i &= 2,600 \\ \sum U_i^2 &= 130,000 \\ s_u &= \$25 \\ \bar{Z} &= \$10 = \text{average income (thousands)} \\ \sum Z_i &= 1,000 \\ \sum Z_i^2 &= 13,600 \\ s_z &= \$6 \text{ (thousands)} \\ \sum U_i Z_i &= 40,000\end{aligned}$$

- a) Make an estimate of total expenditures for camping gear for the 100,000 family units in San Jose by (i) *simple random sampling* and (ii) *ratio estimation*—assume total annual income for all 100,000 units is known to be \$900 million.
- b) Compare the two estimates. Why do they differ? Which is more accurate? Why?
- c) As an alternative, the San Jose area could have been stratified by geographic area into three economic area groups. Estimates of standard deviations of expenditures for camping gear within each area are provided. How would you allocate your sample of 100 items between these groups? What accuracy would you estimate? Compare this to the simple random and ratio estimates above.

Area	Number of Family Units	Estimated Standard Deviation of Camping Expenditures, Dollars
High income.....	30,000	25
Medium income.....	40,000	15
Low income.....	30,000	5
Total.....	100,000	

12. Mr. Worthy, president of Worthy Products, was considering marketing a new product—an ornamental gadget that could be attached to fenders, bumpers, or hoods of automobiles. The gadget would be sold on a door-to-door basis and some automobile owners might buy two, three, or even more.

There were some 200,000 households and some 250,000 automobiles in the territory which Mr. Worthy intended to canvass. In order to make an estimate of his sales in this territory, Mr. Worthy drew a random sample of 50 households and made sales calls. The results of this survey are shown in the table.

RESULTS OF A RANDOM SAMPLE OF 50 HOUSEHOLDS
SURVEY CONDUCTED BY WORTHY PRODUCTS

Household Number	Gadgets Sold	Cars in Household	Household Number	Gadgets Sold	Cars in Household
1	0	0	26	0	0
2	0	2	27	0	2
3	2	4	28	2	4
4	0	1	29	0	1
5	0	0	30	0	0
6	0	0	31	0	0
7	0	0	32	0	0
8	0	2	33	0	2
9	0	2	34	0	2
10	1	3	35	1	3
11	0	1	36	0	1
12	0	1	37	0	1
13	0	1	38	0	1
14	0	2	39	0	2
15	0	3	40	0	3
16	0	2	41	0	2
17	0	0	42	0	0
18	0	1	43	0	1
19	0	1	44	0	1
20	0	2	45	0	2
21	1	3	46	1	3
22	2	3	47	2	3
23	1	1	48	0	1
24	0	2	49	1	2
25	0	1	50	0	1
			Totals	14	76

- a) Treating the sample data as a simple random sample of households, estimate the total sales for all 200,000 households.
 - b) Using the ratio of sales to number of automobiles in a household, estimate total sales.
 - c) Compare the two estimates. Why do they differ? Considering possible bias, which estimate do you think is more accurate?
13. A study was undertaken in a certain city to estimate the total number and types of major appliances (refrigerators, stoves, washers, dryers, dishwashers, freezers). The city was first divided into 600 blocks. From aerial photographs and automobile trips about the city, the number of households in each block was estimated. By this process, it was estimated that there were 10,000 households in the city. Next, 30 blocks were selected at random. In each of these blocks *all* the households were contacted and

information about their appliances was obtained. The results are shown in the table.

Block No.	Number of Appliances	Estimated No. of Households
1	64	16
2	48	14
3	42	5
4	94	20
5	70	13
6	40	11
7	31	12
8	21	6
9	49	12
10	73	22
11	85	23
12	47	17
13	39	8
14	60	14
15	66	20
16	32	8
17	53	12
18	64	24
19	110	27
20	95	28
21	137	40
22	49	9
23	63	15
24	54	15
25	59	11
26	80	19
27	64	17
28	110	24
29	73	26
30	103	33
Totals	1975	521

- a) Estimate the total number of major appliances in the city using the ratio estimate (ratio of number of appliances to number of households in a block).
 - b) Consider the blocks as clusters, with 100 percent second-stage sampling, and make an estimate of total number of major appliances using the cluster sampling approach. Does your estimate differ from that in a? Explain.
 - c) How else might you make an estimate of total number of appliances in the city from the data above?
14. An oil company wanted to estimate the average monthly sales for the next month for its approximately 104,000 credit-card customers. The credit-card accounts were filed by account number in 500 drawers, each containing approximately 200 accounts.
- It was decided first to draw a random sample of 30 drawers and then a

systematic sample of 10 accounts from each drawer selected. The results are shown in the table.

Drawer	No. of Accounts in Drawer	Average Monthly Sales in Sample
1	220	21.67
2	184	19.26
3	200	3.20
4	176	12.17
5	210	5.42
6	208	13.10
7	198	7.15
8	202	10.85
9	206	12.50
10	194	15.47
11	218	17.29
12	217	6.18
13	192	24.53
14	212	8.22
15	202	6.33
16	225	19.13
17	209	7.57
18	208	1.12
19	215	14.71
20	224	6.83
21	216	12.92
22	226	7.21
23	234	34.17
24	196	8.47
25	218	11.16
26	242	9.28
27	200	17.42
28	215	9.64
29	210	22.77
30	204	14.98

- a) Estimate the overall average monthly sales for all 104,620 accounts and the sampling error associated with this estimate.
 - b) What other sampling methods would you suggest that might be more efficient (less sampling error) in this case? How does your method compare with the procedure above in terms of the cost of taking the sample?
15. Consider as a population all the students in your college or department or all the employees in your firm. Determine some variable that you would like to measure for this population, such as the income they expect 10 years after graduation, the average distance they commute to school or work, or the number of hours per week they spend watching television.
 - a) Design a sampling plan to estimate the information you wish. Be sure to define your population exactly. (How do you handle part-time students or employees?) Indicate where you could obtain lists and other information needed for the survey design. Decide upon how accurate you wish the results and how large a sample you will need to achieve this accuracy.
 - b) Prepare a questionnaire to obtain the desired information. Pretest the

questionnaire on a group or groups of persons. Is the survey to be done by mail or personal contact? How will you handle nonresponse?

- c) Conduct the survey and tabulate your results. Estimate the information you desire and determine the sampling error associated with your estimate.
- d) Write up this project in a report form indicating: (i) the sampling plan chosen and why it was chosen, (ii) how the survey was conducted, and (iii) the results of the survey.

SELECTED READINGS

BRYANT, EDWARD C. *Statistical Analysis* (2d ed.) New York: McGraw-Hill, 1966.

Chapter 11 is a concise treatment of various methods of survey sampling.

COCHRAN, WILLIAM G. *Sampling Techniques* (2d ed.) New York: John Wiley, 1963.

This is a textbook on sampling theory and technique. It is at a relatively advanced level and would be useful for students who wish to go into more detail and depth.

CYERT, R. M., and DAVIDSON, N. J. *Statistical Sampling for Accounting Information*. Englewood Cliffs, New Jersey: Prentice-Hall, 1962.

The early chapters deal with the general theory of sampling. Chapter 7 deals with ratio estimation and Chapter 8 with stratified sampling. The treatment in this book is at a moderate level, and examples of sampling in accounting are included.

DEMING, W. EDWARDS. *Sample Design in Business Research*. New York: John Wiley, 1960.

Contains several examples of sampling in business. Replicated sampling is treated in detail. However, the level is advanced and difficult to follow in many places.

HANSEN, M. H.; HURWITZ, W. N.; and MADOW, W. G. *Sample Survey Methods and Theory*. New York: John Wiley, 1953.

Volume I is an authoritative and thorough treatment of sampling methods and applications.

KISH, LESLIE. *Survey Sampling*. New York: John Wiley, 1965.

A modern, comprehensive treatment, incorporating the experience of the University of Michigan Survey Research Center.

SLONIM, MORRIS J. *Sampling in a Nutshell*. New York: Simon and Schuster, 1960.

A light, easily readable treatment of sampling. The whole book (144 pages) can be read in a short time and discusses many important topics, including stratified, cluster, and systematic sampling. Many applications are included.

15. BAYES' THEOREM: REVISING PROBABILITIES

THIS CHAPTER and the next will investigate the process of making decisions based upon information part of which was obtained from a sample. These chapters bring together the elements of decision-making under uncertainty—the subject of Chapters 9 and 10, with the concepts of statistical inference—treated in Chapters 11, 12, and 13. Thus, three factors may contribute to the decision solution: (1) the economic consequences of the various actions; (2) the original probability distribution of the decision-maker; and, now, (3) the added information obtained from a sample. This chapter shows how to revise probabilities in the light of additional information and how to evaluate this information in advance to determine whether we should take a sample—and if so, what size of sample—before acting. Chapter 16 applies this analysis to the case of normal probability distributions.

In Chapter 10, the concept of the expected value of perfect information (EVPI) was introduced. This represented the economic worth, in a given decision situation, of having a perfect predictor of what event would occur. Such a perfect predictor is rarely available. However, it is often possible to take a sample. Any sample estimate has associated with it sampling error and possibly bias, so it is not a perfect predictor. But the sample does give some additional information and should, on the average, improve the decision that is made. Since an improvement in decision-making has an economic value, the sample information has a measurable worth to the decision-maker, and the larger the sample, the greater the value since larger samples are more precise. But larger samples cost more money than smaller samples. And so the problem facing the decision-maker is to pick an optimum sample size that bal-

ances the worth of the sample information with the cost of taking the sample. This sample size might even be zero, meaning that he should act now without sampling. On the other hand, the sample cannot be so large that its cost exceeds EVPI.

A second related question is how the decision-maker should act after he has taken a sample. How much weight should he place upon the sample information relative to his prior probabilities? Should he change his decision because of the sample? There are thus two questions facing the decision-maker in an uncertain situation: (1) Should he take a sample, and if so, how large? (2) Given that a sample has been taken, what action should be taken on the basis of the sample results? Because this second question—the effect of sampling on decision-making—generally is easier to answer than the first, we shall begin with it and return to the first question—on the selection of the sample itself—at the end of the chapter.¹

PRIOR AND POSTERIOR PROBABILITY DISTRIBUTIONS

In order to introduce the concepts of *prior* and *posterior* decision-making or “betting” distributions, let us first consider a rather artificial illustration. Suppose there are two large identical opaque jars on the table in front of you. Each of these jars contains 50 ping-pong balls. Jar A contains all red-colored balls; Jar B contains all white balls. One of the jars is picked by the following random procedure: A fair die is rolled. If a 1 or a 2 turns up, Jar A will be picked; if a 3, 4, 5 or 6 turns up, Jar B will be picked. You are not allowed to witness the rolling of the die. Now, you are asked to play a game in which you guess which jar is to be selected. It is reasonable to assign a probability of $\frac{1}{3}$ to the event “Jar A is picked” since the probability of rolling a 1 or 2 out of six faces on the die is $\frac{1}{3}$. Similarly, the probability of the event “Jar B is picked” is $\frac{2}{3}$. Let us call these our *prior probabilities*. These probabilities represent betting odds about which jar is to be selected.

Now, suppose a jar has been selected (which one you do not know), and you are allowed to take a ball from it and look at it before acting—that is, before guessing “A” or “B.” The drawing of the ball from the jar is essentially taking a sample of size 1. After the sample, what would be your betting odds (called the *posterior* probability distribution) about which jar was selected? It would depend upon the

¹ We consider here taking only a single sample and then acting. This procedure is often desirable, as in making a nationwide business survey involving a large fixed cost. Alternatively, we could take a series of samples and reach a decision whenever the cumulative evidence became convincing one way or the other. Some of these “sequential” sampling plans are described in Chapter 25 on statistical quality control.

color of the ball that was drawn. Since Jar A contains all red balls and Jar B contains all white balls, the color of the ball would give us an errorless indicator of which jar was selected. The betting distributions are shown in Table 15-1.

The important points of this illustration are (1) We have an initial decision-making probability distribution (column 2)—this is designated as the *prior* distribution since it is set up before the sample is taken; (2) this probability distribution is revised after the inclusion of the sample information—this revised distribution is called the *posterior* probability distribution; and (3) the posterior distribution depends upon the sample outcome. There is a different posterior distribution for each sample result.

Table 15-1

PRIOR AND POSTERIOR PROBABILITY DISTRIBUTIONS

Event: Jar Selected Is	Prior Probability (Before Draw)	Posterior Probability	
		If Ball Drawn Is Red	If Ball Drawn Is White
A	0.333	1.0	0.0
B	0.667	0.0	1.0
	1.0	1.0	1.0

Bayes' Theorem

The above example may seem trivial when one jar contained all white balls and the other all red balls. It is not so trivial if we change the problem slightly. Suppose, for example, Jar A contains 70 percent red balls and 30 percent white balls, and Jar B contains 20 percent red balls and 80 percent white balls. Let us see how to determine the posterior probabilities in this case. If only one ball is to be drawn, it can be either red or white. We can draw up the joint probabilities in Table 15-2, as was done in Chapter 7. Recall that a jar (either A or B) was selected at random by rolling the die, and then a ball was selected at random from the designated jar. Hence, we can determine the joint probability of obtaining both a particular jar and a particular color of ball. For example, the joint probability of drawing Jar A and then a red ball is $P(A, R)$ or $P(R, A)$. From page 147, the joint probability can be written as

$$\begin{aligned} P(R, A) &= P(R|A) P(A) \\ &= (0.70)(0.333) = 0.233 \end{aligned}$$

$P(R|A)$ is the conditional probability of a red ball given Jar A; it equals 0.70 since Jar A contains 70 percent red balls. Also $P(A) = 0.333$, the probability of drawing Jar A.

The other joint probabilities in Table 15-2 are computed in a similar manner. The entries at the bottom of the table are the marginal probabilities of obtaining a given color ball. That is, one can obtain a red ball either by drawing Jar A and then a red ball or by drawing Jar B and then a red ball. Thus, the probability of a red ball is the sum of these joint probabilities, that is,

$$P(R) = P(R, A) + P(R, B) = 0.233 + 0.133 = 0.366.$$

We are now ready to revise the prior betting distribution. Suppose that we draw a red ball. We ask this question: What is the probability

Table 15-2
JOINT PROBABILITY TABLE
COLOR OF BALL DRAWN

Jar	Red	White	
A	$P(R, A) = P(R A) P(A)$ $= (0.70)(0.333) = 0.233$	$P(W, A) = P(W A) P(A)$ $= (0.30)(0.333) = 0.100$	$P(A) = 0.333$
B	$P(R, B) = P(R B) P(B)$ $= (0.20)(0.667) = 0.133$	$P(W, B) = P(W B) P(B)$ $= (0.80)(0.667) = 0.534$	$P(B) = 0.667$
	$P(R) = P(R, A) + P(R, B)$ $= 0.233 + 0.133$ $= 0.366$	$P(W) = P(W, A) + P(W, B)$ $= 0.10 + 0.534$ $= 0.634$	1.0

that we have selected Jar A, given the draw of a red ball? Symbolically, we want to find the conditional probability $P(A|R)$. From the definition of conditional probability (Chapter 7),

$$P(A|R) = \frac{P(A, R)}{P(R)} \quad (1)$$

That is, the conditional probability of Jar A, given that a red ball was drawn, is equal to the joint probability of Jar A and a red ball divided by the marginal probability of a red ball. But a red ball may be drawn either from Jar A or B and, hence, the marginal probability may be expressed as the sum of the probabilities of drawing a red ball from Jars A and B. That is,

$$P(R) = P(R, A) + P(R, B)$$

But now the probabilities $P(R, A)$ and $P(R, B)$ may be written as in Table 15-2, column 1:

$$P(R, A) = P(R|A) P(A) \quad \text{and} \quad P(R, B) = P(R|B) P(B)$$

We can then rewrite (1) as

$$P(A|R) = \frac{P(R|A)P(A)}{P(R|A)P(A) + P(R|B)P(B)} \quad (2)$$

Conditional probability expressed in the form of Equation 2 is known as *Bayes' Theorem*.² Note that it expresses the posterior probability of Jar A given a red ball drawn, $P(A|R)$, in terms of the prior probabilities for Jars A and B, $P(A)$ and $P(B)$, and the conditional probabilities of a red ball drawn from Jars A and B [$P(R|A)$ and $P(R|B)$].

Substituting the numerical values in Equation 2, we have

$$P(A|R) = \frac{(0.70)(0.333)}{(0.70)(0.333) + (0.20)(0.667)} = \frac{0.233}{0.366} = 0.637$$

The analogous Bayes' Theorem formula for $P(B|R)$ is

$$\begin{aligned} P(B|R) &= \frac{P(R|B)P(B)}{P(R|A)P(A) + P(R|B)P(B)} \\ &= \frac{(0.20)(0.667)}{(0.70)(0.333) + (0.20)(0.667)} = 0.363 \end{aligned}$$

The values $P(A|R) = 0.637$ and $P(B|R) = 0.363$ are the revised or posterior probabilities that the jar selected is Jar A or Jar B, respectively, given that the sample ball was red. If a white ball had been drawn, then the posterior probabilities could be obtained in a similar manner. They are $P(A|W) = 0.158$ and $P(B|W) = 0.842$.

These posterior probabilities represent "betting odds" in the same

² A more general form of Bayes' Theorem is as follows: Given a set of mutually exclusive and collectively exhaustive events, E_1, E_2, \dots, E_n , and an experimental outcome, e_j ,

$$P(E_j|e_j) = \frac{P(e_j|E_j) P(E_j)}{\sum_{i=1}^n P(e_j|E_i) P(E_i)} \quad \text{for } j = 1, 2, \dots, n$$

sense that the prior probabilities did. There was a $1/3$ chance of Jar A before a ball was drawn. After the draw of a red ball, the chance for Jar A increased to almost $2/3$ (i.e., 0.637); if a white ball is drawn, the odds for Jar A drop to roughly 16 chances in 100 (i.e., 0.158). These results are generally what we would expect from common sense: The draw of a red ball should increase the chances of Jar A since it contains predominantly red balls; and the draw of a white ball should increase the chances of Jar B (and decrease those of A) since it contains predominantly white balls. The use of Bayes' Theorem enables us to attach exact numerical values to the changes in the betting or decision-making probabilities.

Table 15-3

BAYES' THEOREM: COMPUTATION OF POSTERIOR PROBABILITY
(SAMPLE RESULT: ONE RED BALL)

Event: Jar Selected Is	Prior Probability $P(\text{Event})$	Conditional Probability $P(\text{SampleResult} \text{Event})$	Joint Probability $P(\text{Sample Resultand Event})$ (Col. 2 \times Col. 3)	Posterior Probability $P(\text{Event} \text{Sample Result})$ (Col. 4 \div Σ Col. 4)
(1)	(2)	(3)	(4)	(5)
A	0.333	0.7	0.233	$0.233/0.366 = 0.637$
B	0.667	0.2	0.133	$0.133/0.366 = 0.363$
Total	1.000		0.366	1.000
			↑ Marginal Prob- ability = $P(\text{Sam-ple Result})$	

It will be helpful for further analysis to put the computations of the posterior distribution in table form. The general form of the table and the specific calculations which were performed above are repeated in Table 15-3.

Column 1 in Table 15-3 lists the possible events; in this case, Jar A or B. Column 2 shows the prior (i.e., before sample) probabilities: $1/3$ and $2/3$ for Jars A and B, respectively. Column 3 shows the probability of the sample result, given each of the events. In this case it shows the probability of drawing one red ball from Jars A and B, respectively. Column 4 is the joint probability of the event and the sample both occurring. It is obtained by multiplying the values of column 2 times those in column 3.

The sum of the values in column 4 is the marginal probability of the given sample result. In this case, it is the probability of drawing a red

ball, obtained by summing the two probabilities—a red ball drawn from Jar A and a red ball drawn from Jar B.

Column 5 shows the posterior probabilities, obtained by dividing the individual column 4 values by the column 4 total. The total of column 4 is the probability of a red ball, but since the red ball in fact has been drawn, its probability must be “blown up” to 1.00. The other values in column 4, therefore, are “blown up” or increased in the same proportion, giving the column 5 posterior probabilities.

Revision of Probabilities: Binomial Sampling

Let us continue the above illustration one more step. Suppose that we were to draw a sample of 3 balls from the unidentified jar that was

Table 15-4
CALCULATION OF POSTERIOR PROBABILITIES
(SAMPLE OF 2 RED BALLS AND 1 WHITE BALL)

Event: Jar Selected Is	Prior Probability	Conditional Probability $P(r = 2 n = 3, p)$	Joint Probability (Col. 2 \times Col. 3)	Posterior Probability (Col. 4 $\div \Sigma$ Col. 4)
(1)	(2)	(3)	(4)	(5)
A (with $p = 0.7$)	0.333	0.441	0.147	$0.147/0.211 = 0.697$
B (with $p = 0.2$)	0.667	0.096	0.064	$0.064/0.211 = 0.303$
	1.000		0.211	1.000
			↑ Marginal Probability of This Sample	

selected (replacing each after it is drawn). Further suppose that of the three balls, two were red and one was white. How would we obtain the posterior probabilities? First let us ask how we can obtain the conditional probabilities for this sample (2 red, 1 white), that is, $P(\text{sample} | \text{Jar A})$ and $P(\text{sample} | \text{Jar B})$. Since Jar A contains 70 percent red balls, the probability of a sample containing 2 red balls and 1 white ball is simply the binomial probability $P(r = 2 | n = 3, p = 0.7) = 0.441$ (from Appendix F). Similarly, the probability of the sample given Jar B (with 20 percent red balls) is the binomial probability $P(r = 2 | n = 3, p = 0.2) = 0.096$. With these numbers we can fill in the remainder of Table 15-4 to determine the posterior probabilities.

It is important to understand that both the prior and posterior distributions are betting distributions. Before any sample information, we

would bet on Jar B with odds of 2 out of 3. After this sample, the odds change considerably in favor of Jar A (to 0.697 probability).

In Table 15-4, the sum of column 4 is 0.211. This is the probability of obtaining this particular sample (2 red, 1 white) when drawing three balls. Other possible sample results are shown in Table 15-5.

Thus the marginal probability of obtaining a sample with three red balls is 0.120. And if this sample were to occur, the posterior probabilities would be 0.950 for Jar A and 0.050 for Jar B. The calculations of the results shown in Table 15-5 are not shown, but the numbers can be obtained by setting up a table, such as Table 15-4, for each possible sample result.

Table 15-5

POSSIBLE SAMPLES OF SIZE THREE
AND POSTERIOR DISTRIBUTIONS

Sample Result	Marginal Probability	Posterior Probability of	
		Jar A	Jar B
3 red balls	0.120	0.950	0.050
2 red, 1 white	0.211	0.697	0.303
1 red, 2 white	0.319	0.197	0.803
3 white	0.350	0.026	0.974
Total	1.000		

POSTERIOR PROBABILITIES AND DECISION-MAKING

The discussion above has concentrated upon the revision of probabilities and neglected the economic information in the decision process. Let us reintroduce the economic payoffs by means of an example. A manufacturer of electronic equipment operates two factories: one that manufactures components and another that assembles the components into complete units. A certain part is shipped from the manufacturing plant to the assembly plant in lots of 5,000 units. It has been very difficult to regulate the quality of this particular part; lots have been received with as little as 1 percent of the parts defective to as high as 20 percent of the parts defective. The fraction defective p (i.e., percent divided by 100) in the last 20 lots received is shown in Table 15-6. Let us suppose that management is willing to use these historical frequencies as a betting distribution about the fraction defective in the next lot.³

³ Perhaps a more reasonable procedure would call for smoothing this frequency distribution to give some probability to the intermediate values of p . For a procedure to do this, see Chapter 4, page 84.

Table 15-6
FRACTION DEFECTIVE FOR LOTS
OF THE SPECIFIED PART

Fraction Defective (p)	Number of Lots with this Fraction Defective	Relative Frequency
0.01	3	0.15
0.02	5	0.25
0.05	7	0.35
0.08	3	0.15
0.10	1	0.05
0.20	1	0.05
Totals	20	1.00

Economic Analysis before Sampling

When a defective part goes unnoticed and is assembled into the final unit, it affects the performance of the final unit. In such cases, the final unit has to be torn down and the defective part replaced. The cost of this tearing down and reassembling a final unit is \$1.50 each.

An alternative is to inspect the entire incoming lot of parts and to remove all defective parts before assembly. The cost of this 100 percent inspection is 10 cents per part or \$500 per lot. A lot of the particular part has just arrived and the manager must decide whether to inspect 100 percent or to use the lot as is. Let us first draw up a payoff table for this decision problem. This is done in Table 15-7.

Table 15-7
PAYOFF TABLE FOR ACTIONS "INSPECT 100 PERCENT" AND "ACCEPT LOT AS IS"
(Lot Size 5,000; Inspection Cost 10 CENTS; Replacement Cost \$1.50)

Event: Fraction Defective in the Lot (p) (1)	Probability $P(p)$ (2)	Costs*		Opportunity Losses	
		Inspect 100 Percent (3)	Accept Lot as Is (4)	Inspect 100 Percent (5)	Accept Lot as Is (6)
0.01	0.15	\$500	\$ 75	\$425	\$ 0
0.02	0.25	500	150	350	0
0.05	0.35	500	375	125	0
0.08	0.15	500	600	0	100
0.10	0.05	500	750	0	250
0.20	0.05	500	1,500	0	1,000
Expected Values		\$500	\$ 382.50	\$195	\$ 77.50

* Note that we have linear cost equations in this example. Cost of inspection = \$500. Cost of accepting as is = $(\$1.50)(5,000)p$, where p is the unknown variable (fraction defective). $E(p)$ can be calculated to be 0.051 and, hence, the expected cost can be determined as $E(c) = (\$1.50)(5,000)E(p) = \$7,500(0.051) = \$382.50$, as above.

Columns 1 and 2 come from Table 15-6. Costs in columns 3 and 4 are determined as follows: for 100 percent inspection, cost is 10 cents per unit times 5,000 parts = \$500; for accepting the lot as is, the cost is \$1.50 per unit replacement cost times the number defective ($5,000 \times p$). For example, when $p = 0.05$, we expect $0.05 \times 5,000 = 250$ defectives, and $250 \times \$1.50 = \375 . Opportunity losses in columns 5 and 6 are obtained by subtracting the lower of two costs in each row from the costs themselves. Expected values are the weighted averages of the figures in each column multiplied by their probabilities and totaled.

As can be seen from this table, the best action is to accept the lot as is, since this action has the lower expected cost, even though this will necessitate some rework at a later time. The EVPI is \$77.50 per lot (the expected opportunity loss of the best alternative). Since this is a fairly substantial amount, the decision-maker should investigate ways of obtaining additional information.

Economic Analysis after Sampling

One method of obtaining at least partial information in this situation is by taking a random sample of parts in the lot and inspecting the items in the sample. From the number of defects in the sample we can make some inferences about the fraction defective in the entire lot.

Let us suppose that the manager arbitrarily decided to sample 25 items from the lot and that he found that two of 25 were defective. We now want to investigate what action should be taken on the basis of his prior probabilities and the sample information combined. The decision-maker can revise his original or prior betting distribution in the same fashion as in Table 15-4. This is done in Table 15-8.

Compare the posterior probabilities with the prior probabilities. The fraction defective in the sample was $2/25 = 0.08$. Note that the posterior probabilities for values of p close to 0.08 have increased (relative to the prior values) and the posterior probabilities for p far from 0.08 have decreased.

We can now use the posterior probabilities, together with the original costs in Table 15-7 to revise our payoff table, using the same computations as before.⁴ (See Table 15-9.) The optimal action remains to accept the lot as it is, since this action has the lower expected cost. However, the expected cost is somewhat more than previously, since the

⁴ We can find the $E(p)$ for the posterior distribution = 0.0609. As an alternate method of finding the expected cost, we have $E(c) = (\$1.50)(5,000)E(p) = \$7,500 \times (0.0609) = \456.75 as in Table 15-9.

Table 15-8

CALCULATION OF POSTERIOR PROBABILITIES BY BAYES' THEOREM
(SAMPLE OF 25 PARTS, WITH 2 DEFECTIVES)

Event: Lot Fraction De- fective Is p (1)	Prior Probability $P(p)$ (2)	Conditional Probability * $P(r = 2 n =$ $25, p)$ (3)	Joint Probability $P(p)P(r = 2 n =$ $25, p)$ (Col. 2 \times Col. 3) (4)	Posterior Probability $P(p)P(r = 2 n = 25, p)$ $\Sigma P(p)P(r = 2 n = 25, p)$ (Col. 4 $\div \Sigma$ Col. 4) (5)
0.01	0.15	0.024	0.00360	0.002
0.02	0.25	0.075	0.01875	0.115
0.05	0.35	0.231	0.08085	0.498
0.08	0.15	0.282	0.04230	0.261
0.10	0.05	0.266	0.01330	0.082
0.20	0.05	0.071	0.00355	0.022
	1.00		0.16235	1.000
			↑ Marginal Probability of This Sample	

* The values in column 3 were obtained from the Binomial Tables, Appendix F.

fraction defective in the sample (0.08) exceeded the expected fraction defective (0.051) prior to taking the sample (Table 15-7, footnote). Note that the posterior EVPI is still quite large (\$68.60 from Table 15-9), indicating that the particular sample result did little to resolve the uncertainty about which action to take. The decision-maker could consider taking a second sample before acting.

The sample result "2 defectives out of 25" is only one of many that

Table 15-9

PAYOFF TABLE USING POSTERIOR PROBABILITIES
(SAMPLE OF 25 PARTS WITH 2 DEFECTIVES)

Event: Fraction Defective in the Lot p	Posterior Probability $P(p)$	Costs		Opportunity Losses	
		Inspect 100 Percent	Accept Lot as Is	Inspect 100 Percent	Accept Lot as Is
0.01	0.022	\$500	\$ 75	\$425	\$ 0
0.02	0.115	500	150	350	0
0.05	0.498	500	375	125	0
0.08	0.261	500	600	0	100
0.10	0.082	500	750	0	250
0.20	0.022	500	1,500	0	1,000
Expected Values		\$500	\$ 456.75	\$111.85	\$ 68.60

could have occurred. The other possible results are shown in Table 15-10. The decision action changes if 3 or more defectives are found in the sample—then 100 percent inspection become the more economical decision. Note that different sample results lead to quite different values of the posterior EOL of the better action, or EVPI. When either very few or very many defectives are found in the sample the decision to be taken becomes relatively clear. When a “middle” number of defectives is found (around 2 or 3 out of 25), there remains considerable uncertainty about which is the correct action. This is true of sampling in general. Very good or very bad sample results lead to clear-cut deci-

Table 15-10
POSSIBLE RESULTS FOR A SAMPLE OF 25 ITEMS

Sample Result (Number of Defectives) <i>r</i>	Posterior Action	Posterior Expected Cost	Posterior Expected Opportunity Loss
0	Accept as is	\$212.25	\$ 8.05
1	Accept as is	333.22	26.95
2	Accept as is	456.75	68.60
3	Inspect	500.00	63.92
4	Inspect	500.00	32.55
5	Inspect	500.00	13.00
6	Inspect	500.00	4.38
7 or more	Inspect	500.00	Very small

sions; borderline results are indecisive and may require further sampling. If you sampled a half-dozen apples out of a bushel basket full, and all were good, you might readily accept the basket, but if one were bad, you would be uncertain.

EXPECTED VALUE OF SAMPLE INFORMATION

In the previous section, we addressed ourselves to the question, “Given that a sample of a certain size has been drawn, what action should be taken on the basis of both prior and sample information?” In this section we examine the question, “Should we take a sample, and if so, how large should it be?” As noted earlier, sampling may be costly, and the larger the sample, the greater the cost. Hence, to take a sample, we must determine that the economic value of the information contained in the sample is worth the cost.

A sample has value because it reduces the uncertainty of the decision situation. After the sample, we are more sure than before about which

event will occur. Hence, we are less apt to make a costly mistake. To see this, compare the EVPI prior to taking the sample which is \$77.50 (Table 15-7) with the posterior expected opportunity losses (or EVPI's) in Table 15-10. After the sample, the EVPI ranges from near 0 (when $r = 7$ or more) to a high of \$68.60 (when $r = 2$). All the values are below \$77.50, indicating that even the most inconclusive sample result ($r = 2$) somewhat reduces the uncertainty. And the sample result ($r = 0$) has a posterior EVPI of \$8.05, a considerable reduction. Thus, a sample result "0 defectives out of 25" makes it almost certain that the correct action is to accept the lot as is. In this case the sample information is quite conclusive.

Another way of determining the value of a given size sample before taking the sample is to compare the expected cost (or profit) before the sample with the expected cost (or profit) if we had taken the sample. The amount by which cost is reduced from the before-sample case to after-sample case gives us the economic value of the sample. The prior expected cost is determined, in our example, as \$382.50 from Table 15-7. The posterior expected cost, however, depends upon the particular sample result that might occur. For example, the posterior expected cost would be \$456.75 for a sample result of 2 defectives out of 25 (see Table 15-9). Similar expected cost values can be calculated from the posterior distributions associated with other sample results. These calculations are not shown, but the results are displayed in Table 15-10. The lowest posterior expected cost would be \$212.25, if zero defectives were observed in the sample. At the other extreme, if 3 or more defectives were observed, 100 percent inspection is the action chosen with a certain cost of \$500.

How can we compare prior with posterior expected cost if posterior expected cost is represented by several possible values? The answer lies in the use of an average or expectation of the posterior costs. Recall that we can determine the marginal probability of any particular sample result for a given set of prior probabilities. Thus, the probability of exactly 2 defectives out of 25 items is found in Table 15-8 (sum of column 4) to be 0.162. Similarly, the probability for the sample result "0 defectives out of 25 items" can be found to be 0.387 (calculations are not shown); the probability for the sample "1 defective out of 25 items" is 0.286; and so on, as shown in column 2 of Table 15-11.

These probabilities can be used as weights to determine the expectation or average of the posterior expected costs associated with each possible sample result. This calculation is performed in Table 15-11.

The amount of \$333.93 from Table 15-11 is our expectation, before

taking the sample, of what the posterior expected cost will be. The value of the sample, called *expected value of sample information* or EVSI, is the difference between the prior expected cost (\$382.50) and this value. It is $\$382.50 - \$333.93 = \$48.57$. This is the amount by which we can expect to reduce cost by taking a sample of 25 items and then acting on the basis of the sample result. If the cost of taking the sample of 25 items is less than \$48.57, therefore, the sample should be taken. In our example, inspection cost is only 10 cents a part, or \$2.50 for 25 parts, so the sample would be worthwhile.

Table 15-11

ESTIMATING POSTERIOR EXPECTED COST, BEFORE SAMPLING

Sample Result (Number of Defectives) r (1)	Probability of Sample Result $P(r)$ (2)	Posterior Expected Cost (3)	Expected Value (Column 2 \times Column 3) (4)
0	0.387	\$212.25	\$ 82.14
1	0.286	333.22	95.30
2	0.162	456.75	73.99
3	0.082	500.00	41.00
4	0.039	500.00	19.50
5	0.020	500.00	10.00
6	0.011	500.00	5.50
7 or more	0.013	500.00	6.50
	1.000		\$333.93

Note that the expected value of sample information is a value obtained before the sample has been taken—in fact, before the decision has been made about whether a sample ought to be taken at all. It is an *expected* value. Before sampling we do not know how much the sample will save; we do not know even what the sample result will be and, hence, are uncertain what action we will take based upon the sample result. Using the probabilities of the various sample results and computing the expected value, we are determining the “best bet” to make in the decision situation.

Throughout this example we have examined only the possibility of a sample of 25 items. Would not a sample of 20 items or 50 items or 100 items be better? The low inspection cost (10 cents per part versus \$1.50 replacement cost) and the initial uncertainty as to fraction defective (as shown by the diffuse probability distribution in Table 15-7) suggest that the optimum sample size should be much larger than 25. On the

other hand, it would not pay to take a sample so large that its cost exceeded the expected value of perfect information, which was \$77.50 (page 364). Hence, the sample size should not exceed 775 (since $\$77.50 \div 0.10 = 775$), out of the whole lot of 5,000 parts. We could then take a few sample sizes—say, from 50 to 700—and compute EVSI less sampling cost for each to determine the optimum sample size. These calculations would be tedious and might be more costly to perform than the savings from taking a sample were it not for the availability of electronic computers.⁵

Fortunately, we have techniques for the special case of normal sampling (or the normal approximation to the binomial in this case) that reduce all this computation to a single formula. However, because it is necessary to understand the concept of the expected value of sample information (EVSI) and how it can be obtained in a general case, we have gone through the detailed procedure above. The special case will be the subject of the following chapter.

BAYESIAN VERSUS CLASSICAL APPROACH

There is some controversy in the statistics profession over the validity of the decision-making approach suggested in this chapter. Our approach is in accord with the thinking of the Bayesian school. The more traditional or “classical” approach to the evaluation of sample information was presented in Chapters 11 to 13. The controversy centers about whether the statistician, as a scientist, should be concerned only with the objective evidence of the sample (classical school) or whether he should also be concerned with the whole decision framework, including any subjective judgment of the decision-maker about the probabilities of various events. Bayesian analysis takes into account subjective probabilities and utility values in much the same way as they are intuitively considered by the business executive.

A prior judgment is particularly significant if sample information is meager, as in most small samples. In taking very large samples, where the evidence is overwhelming, the prior judgment well may be discarded. How much additional information is needed for its evidence to “swamp” prior probabilities? Bayes’ formula provides an answer in the form of an automatic adjustment: If the sample is small, its results may modify prior probabilities but little; but as the sample increases in size,

⁵ J. Pratt, H. Raiffa, and R. Schlaifer in *Introduction to Statistical Decision Theory* (New York: McGraw-Hill, 1965), p. 5, report that computer programs are now being developed for these and other related calculations.

the posterior probabilities approach those shown in the sample, irrespective of the prior judgment.

Bayesian methods also take into account the economic profits or losses of decisions, as well as the probabilities involved. Thus, in the classical testing of hypothesis discussed in Chapter 12, we reject a hypothesis if the risk of making a Type I error—rejecting a true hypothesis—exceeds some critical probability such as 5 percent. This figure is rather arbitrary, and it does not provide for balancing the relative cost of Type I versus Type 2 errors. It is difficult to balance these errors in classical theory. Bayesian statistics adds the economic dimension to the decision-making process and offers an objective criterion for making decisions: Set up a probability distribution and payoff table, then maximize expected profits.

The Bayesian approach thus serves as the completion of the classical theory of statistical inference, through providing the decision-maker with a logical framework within which to apply both his judgment and sample evidence, in proper proportions, to the economic consequences of his possible actions.

SUMMARY

The subject of this chapter is the application of Bayes' Theorem to decision-making under uncertainty. This involves the combination of a *prior* probability distribution (which may be subjective) with the results of a sample to form a *posterior* decision-making distribution.

Bayes' Theorem is a form of expressing the conditional probability of an event, given a sample outcome, in terms of the prior probability of the event and the conditional probabilities of sample result, given the event. Thus, in our first example, the conditional probability of selecting Jar A, given that a red ball has been drawn, is

$$P(A|R) = \frac{P(R|A)P(A)}{P(R|A)P(A) + P(R|B)P(B)}$$

In the electronic component example, we are given prior probabilities for various levels of fraction defective, but if we then take a sample of 25 and find 2 defectives, we can modify the priors by the sample result, as in Table 15-8, to find the posterior probabilities. These revised probabilities are then used in a payoff table, just as the prior probabilities were, to find the expected cost (or profit) of each possible action. In our example, the best decision before sampling was to accept

the lot as is rather than inspect 100 percent. After taking a sample of 25, however, we arrived at a better decision rule: Accept the lot if the sample has 2 or less defectives; otherwise, inspect 100 percent. Each possible sample result has a different posterior distribution and a different posterior expected value.

A sample has economic value because it reduces the uncertainty associated with decision-making. The specific value, called the *expected value of sample information*, is determined by subtracting the expected cost posterior to the sample from the prior expected cost. The expected posterior cost is obtained as an expectation or average of the expected costs associated with the various possible sample results. We can determine if a sample of a given size should be taken at all by comparing the cost of the sample with the expected value of sample information. An optimal sample size can be determined by making this comparison for several sizes of samples, from zero up to the sample size whose cost equals EVPI.

PROBLEMS

1. Explain:
 - a) Prior and posterior distributions.
 - b) Bayes' Theorem.
 - c) Conditional, joint probabilities.
 - d) Posterior expected cost.
 - e) Expected value of sample information.
 2. For the example used in the text on page 360, verify the posterior probabilities $P(A|W) = 0.158$ and $P(B|W) = 0.842$.
 3. Verify the posterior probabilities shown in Table 15-5.
 4. Verify the calculations shown in Table 15-10 for the row listed below, as assigned:
 - a) The row for 0 defectives.
 - b) The row for 1 defective.
 - c) The row for 3 defectives.
 - d) The row for 4 defectives.
 5. In certain portfolio, 70 per cent of the industrial stocks increased in value over the past year while 40 percent of the utility stocks increased. The portfolio contains 80 percent industrial stocks.
 - a) If a stock is selected at random, what is the probability that it is one that has increased in value?
 - b) Suppose a stock is drawn and noted to be one that has increased. What is the probability that the stock is an industrial stock?
-

6. Of the firms in a certain industry, the median age of the chief executive officer is 50 years. Of those executives under 50, 65 percent were in marketing before becoming president. Of those over 50, only 45 percent reached the chief executive position through marketing.

If a chief executive is selected at random in this industry, and if it is noted that he had *not* reached the top through marketing, what is the probability that he is over 50 years old?

7. The Glorious Eastern Motel Association is about to poll its members about whether or not to accept a certain national credit card. The executive secretary of the association feels that he knows "pretty well" how many (i.e. what percent) of the motels favor accepting the credit card. Suppose he attaches the following probabilities to various percents in favor:

Percent of Motels in Favor of Ac- cepting the Credit Card	Probability of Exactly that Percent
30	0.10
40	0.30
50	0.40
60	0.20
	1.00

- a) Based on this information, would you guess that a vote for the credit card would win or lose?
- b) Suppose you drew a random sample of 15 motels and find 8 in favor and 7 opposed. What probabilities would you then assign to "percent of motels favoring accepting the credit card"?
- c) After the above sample, what is the probability that a vote will find a majority in favor?
8. The director of another association of motels wanted to know the feelings of the majority of the membership on a matter of policy. The director had only vague notions about the opinions of the members on this issue; however, he was able to draw up the following prior distribution:

Event: Proportion of Members in Favor of New Policy	Prior Probability
20	0.05
30	0.10
40	0.20
50	0.30
60	0.20
70	0.10
80	0.05

A sample of 25 members were selected at random and their opinion was obtained with the following result: 10 members were in favor of the new policy and 15 were opposed. The director considered this conclusive evidence that the new policy was not favored by a majority of the members. Do you agree with this conclusion?

9. An election is being held in a certain plant to determine if the workers should be represented by a union. A few days before the election, management assigns the probabilities below to the events, "Proportion of workers who will vote for unionization":

Event: Proportion of Workers Voting for Union	Probability
0.35	0.15
0.40	0.30
0.45	0.20
0.50	0.20
0.55	0.10
0.60	0.05
	1.00

A sample of 20 workers is chosen at random and the voting intentions of each ascertained with the following results:

11 will vote for unionization
 9 will vote against unionization
 20 total

After the sample, what probabilities should management assign to the events "Proportion of workers voting for union"?

10. From past experience, the fraction of items defective in lots manufactured by a certain process has the following distribution:

Event: Lot Fraction Defective	Relative Frequency
0.01	0.50
0.02	0.30
0.05	0.10
0.10	0.05
0.15	0.05
	1.00

A sample of 15 items are taken from a certain lot and no defectives are found. What posterior probabilities would you assign to the event "Lot fraction defective"?

11. The Theta Company manufactures its requirements for part No. 805 in lots of 1,000 units. It has been difficult to control the quality of this product without a complicated readjustment of the manufacturing equipment. The cost of such a readjustment is \$400. When such a readjustment has been made, only 2 percent defectives are produced. Without the adjustment, the quality has been quite variable, as shown by the history of the last 20 lots.

Fraction Defective Without Adjustment	No. of Lots
0.02	5
0.05	8
0.10	4
0.15	2
0.20	1
	<u>20</u>

A lot of part No. 805 is about to be manufactured, and management is undecided about whether it should pay for the costly adjustment or take the chance of a large percent of defectives. Defective items cost \$5 each in replacement cost.

- a) Draw up a payoff table and calculate the expected cost of each action, using the past frequency data as prior probabilities. Which action is preferable?
 - b) What is the EVPI?
 - c) Suppose the manufacturing process was set up and the first 20 items were examined and 2 defectives were found. Should the machine be shut down and an adjustment made at this time or should the manufacturing process be allowed to continue?
12. (Continuation of Problem 11.) Suppose the sample result had been 0 defectives out of 20 items sampled. What is the expected posterior cost of each action? Which action is preferable? What is the posterior EVPI?
13. (Continuation of Problems 11 and 12.)
- a) Find the expected posterior cost for other relevant sample results.
 - b) What is the expected value of sample information for a sample of 20 items in this decision situation?
 - c) Suppose it cost \$20, plus \$2 per item sampled. Should a sample of 20 items be taken?
14. As president of the Alma Mater University Alumni you are planning the annual alumni banquet. There are 1,000 members of the alumni chapter. Based upon the attendance of previous years, you assign the following probabilities to the number attending this years' annual banquet:

No. At- tending	Probability
100	0.2
200	0.2
300	0.3
400	0.2
500	0.1

The banquet is to be held at the Ritz-Oasis, and the banquet manager informs you that you must specify the number you expect to attend within the next few days. He gives you a price of \$6 per plate for the exact number specified. Additional dinners (beyond the number specified) may be obtained on the day of the banquet (after registration when exact attendance is known) at a price of \$8 each. If fewer dinners are needed than ordered, a partial refund of \$2 will be made for each dinner not needed (i.e., \$4 will be charged for each dinner ordered that is not needed).

The fee that you will charge the alumni has been set at \$10 each for those attending. Because of the short time available it is not possible to use a mail reservation system.

- a) Based only on the information given above, how many dinners should you order? What is the EVPI? (Only consider ordering dinners in even hundreds.)
- b) Suppose that you select a random sample of 20 alumni and call them on the phone. Eight indicate that they will attend. Using this sample information, and that above, what action (number of dinners to order) would you take? What is your EVPI?

SELECTED READINGS

Selected readings for this chapter are included in the list which appears on page 396.

16. BAYES' THEOREM FOR NORMAL DISTRIBUTIONS

IN THE PREVIOUS CHAPTER we considered the general case of Bayes' Theorem and its application to decision-making. This chapter will consider a special case, with specific assumptions about (1) the shape of the prior decision-making distribution, (2) the distribution of sample means, and (3) the form of the opportunity loss functions. These assumptions will be explained in detail as they are introduced. They enable us to express Bayes' Theorem and the economic evaluation of sampling in simple formulas. Although the chapter deals with a specialized situation, it is a situation that has wide practical applicability.

DETERMINING THE POSTERIOR DISTRIBUTION

The posterior decision-making distribution results from combining the sample information with the prior probabilities of the decision-maker.

The Distributions Involved

Since there are several distributions or populations involved in the analysis, we shall summarize them below, together with the symbols used.

1. *The Population from Which the Sample Is to Be Drawn.* The population from which the sample is to be drawn is a collection of elements in the real world (people, houses, accounts etc.) which can be classified by some characteristic (income, number of rooms, dollars outstanding, etc.). By taking a sample of these elements, the decision-maker can obtain some information which will help him make his decision. In particular, the sample mean \bar{X} gives an estimate of μ , the unknown mean of the population.

	<i>Random Variable</i>	<i>Mean</i>	<i>Standard Deviation*</i>
1. Population from which sample is drawn (can be any type of distribution)	X	μ	σ
2. Prior distribution of the population mean (assumed normal)	μ	M_0	S_0
3. Distribution of the sample mean (normal for large samples)	\bar{X}	μ	$\sigma_{\bar{X}}$
4. Posterior distribution of the population mean	μ	M_1	S_1

* σ is generally unknown but can be estimated from sample value: $s \approx \sigma$. The $\sigma_{\bar{X}}$ is the standard error of the mean which can also be estimated from a sample: $s_{\bar{X}} \approx \sigma_{\bar{X}}$.

This population distribution can be of any shape. It will often be skewed to the right in economic phenomena. Like the mean μ , the standard deviation σ is also generally unknown, and is usually estimated from the sample data. For large samples, the use of the sample value s in place of σ causes little error.

2. The Prior Distribution. The prior decision-making distribution is a betting distribution representing the decision-maker's uncertainty about the unknown value of the mean μ of the population to be sampled. The mean of this prior distribution M_0 is the decision-maker's best guess of μ . And the standard deviation S_0 is a measure of his uncertainty about μ . If the decision-maker were quite uncertain and believed that μ can have any of a wide range of values, he would make S_0 large. On the other hand, if he felt that μ lay within a narrower range, he would make S_0 small.

Note that the standard deviation of the prior distribution S_0 is *not* an estimate of the standard deviation σ of the population to be sampled. Such an estimate of σ would often be needed, but it is not at all related to the estimates for the prior distribution. To repeat, S_0 is a measure of the decision-maker's uncertain knowledge only about μ , the mean of the population to be sampled.

Assumption (1): The Prior Distribution Is Normal. The use of a normal decision-making distribution is quite appropriate in many situations.¹ The normal distribution is symmetric, implying that the decision-maker's guess of μ is as likely to be off a given amount in either

¹ The use of normal distributions in decision-making was discussed on pages 234 to 241. The reader may wish to review these pages before proceeding.

direction about M_0 . The normal distribution has probability clustered close to M_0 , indicating that the decision-maker's guess is more likely to be close to the true μ than to be far away, and using the normal distribution implies betting odds of roughly 2 out of 3 that μ lies in the range $M_0 \pm S_0$ and odds of about 95 out of 100 that μ is in the $M_0 \pm 2S_0$ range.

3. The Distribution of Sample Means. The sample mean, \bar{X} , is used to estimate the mean μ of the population to be sampled. The sampling distribution of \bar{X} is a theoretical distribution consisting of all possible sample means of a given size drawn from the first population above.²

Assumption (2): The Sampling Distribution of \bar{X} Is Normal. This is not a very restrictive assumption. From the Central Limit Theorem we know that for moderate to large samples the distribution of the sample mean \bar{X} is approximately normal with mean μ (the population mean) and standard deviation $\sigma_{\bar{X}}$, where $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. The value $\sigma_{\bar{X}}$ is a measure of sampling error of \bar{X} . When $\sigma_{\bar{X}}$ is small, the sample contains relatively precise information about μ ; when $\sigma_{\bar{X}}$ is large, the sample information gives a more diffuse estimate of μ .

When the standard deviation of the population σ is estimated by s , the standard error of the sample mean is calculated as $s_{\bar{X}} = s/\sqrt{n}$.

4. The Posterior Distribution. The posterior distribution, like the prior distribution, is a decision-making or betting distribution. It represents the decision-maker's uncertainty about the unknown value of μ after taking into account sample evidence. *If the prior distribution and the distribution of sample means are both normal, then the posterior distribution is also normal.*³ That is, if assumptions (1) and (2) above are satisfied, the posterior distribution is normal. Its mean M_1 and standard deviation S_1 are determined as follows:

$$M_1 = \frac{\frac{M_0}{S_0^2} + \frac{\bar{X}}{\sigma_{\bar{X}}^2}}{\frac{1}{S_0^2} + \frac{1}{\sigma_{\bar{X}}^2}} \quad (1)$$

² See pages 254 to 259.

³ Actually, the normality of the posterior distribution is rather insensitive to violations in the normality of the prior distribution. Schlaifer makes the following statement:

"If the variance of the decision-maker's true prior distribution is large compared with the sampling variance of \bar{X} , he can simplify his calculations with no material loss of accuracy by substituting the mean and variance of his true prior distribution into the formulas which apply to a normal prior distribution."

See R. Schlaifer, *Introduction to Statistics for Business Decisions* (New York: McGraw-Hill 1961), p. 309.

and

$$\frac{1}{S_1^2} = \frac{1}{S_0^2} + \frac{1}{\sigma_{\bar{x}}^2} \quad (\text{the denominator in Formula 1}) \quad (2)$$

Note that:

a) The posterior mean is a *weighted average* of the prior mean and the sample mean, the weights being the reciprocals of the variances of the two distributions. A smaller variance means a higher precision of the mean and hence a greater weight. Thus, if the prior distribution is relatively narrow (i.e., S_0 is smaller than $\sigma_{\bar{x}}$ and hence $1/S_0^2$ is larger than $1/\sigma_{\bar{x}}^2$), the prior mean receives greater weight. But if the sample is relatively precise (i.e., $\sigma_{\bar{x}}$ is smaller than S_0 , and hence $1/\sigma_{\bar{x}}^2$ is greater than $1/S_0^2$), the sample mean receives greater weight. If there were little prior knowledge, the prior standard deviation S_0 would be very large, and the posterior distribution would reflect almost entirely the sample result.

b) The weight received by the sample mean depends upon n , the size of the sample. Recall that $\sigma_{\bar{x}} = \sigma/\sqrt{N}$. As n increases, $\sigma_{\bar{x}}$ decreases, and the sample becomes more precise. Thus, as sample size increases, the weight received by the sample mean ($1/\sigma_{\bar{x}}^2$) increases, and the posterior distribution is more influenced by the sample result. For very large samples, the prior distribution is "swamped out" and has virtually no effect upon the posterior distribution.

c) The reciprocal of the posterior variance is the sum of the reciprocals of the variance of the prior and the sampling distributions.⁴ This implies that the posterior variance (or standard deviation) is smaller than either the prior or posterior variance (or standard deviation). In other words, there is less uncertainty in the posterior distribution than in either of the others.

Assumption (3): Two-Action Problem with Linear Profit Functions. Assumptions (1) and (2) above are enough to guarantee that the posterior distribution is normal. This result may be sufficient to deal with certain decision situations. However, we shall introduce an additional assumption, as we did in Chapter 10. We shall restrict the analysis to problems in which there are only two actions, and the profits (or costs) for each action may be represented by a linear function. This assumption will enable us to reduce the calculation of the expected

⁴ For further discussion, see R. Schlaifer, *Introduction to Statistics for Business Decisions*, p. 302 f.

profit, the expected value of perfect information, and the expected value of sample information, to simple formulas.

An Example

A wholesale merchant has an opportunity to buy a special lot of merchandise for \$10,000. The lot contains 100,000 novelty items at a unit cost of 10 cents, which the wholesaler could sell in turn to his customers for 20 cents each. The wholesaler did not think he could sell all 100,000 items but noted that he had only to sell 50,000 to break even. His prior judgment was that he would sell 54,000, but there was some uncertainty about this sales level. The wholesaler expressed his uncertainty about sales in the form of a normal distribution with mean 54,000 units and standard deviation 10,000 units. This meant that the wholesaler would be willing to bet, with even odds, that sales would be above (or below) 54,000, and he would be willing to give 2 to 1 odds that sales would be in 44,000 to 64,000 range ($54,000 \pm 10,000$). Such odds reflected his experience with similar merchandise.

The wholesaler has 2,000 customers who regularly buy from him. Let us express these preliminary estimates in terms of sales per customer by dividing the above estimates by 2,000. Thus, the prior mean is $M_0 = 54,000/2,000 = 27$ and the prior standard deviation is $S_0 = 10,000/2,000 = 5$. In these terms, the decision-maker's best guess (M_0) is that he will sell an average of 27 units per customer, and the standard deviation about this guess (S_0) is 5 units per customer. The break-even level of sales (K) is an average of 25 units per customer.

We can express the profit equations as follows:

$$\begin{aligned}\text{Profit for action "Buy the lot": } \pi &= -10,000 + (0.20)(2,000)\mu \\ &= -10,000 + 400\mu \text{ in dollars} \\ \text{Profit for action "Do not buy"} \pi &= 0\end{aligned}$$

In the first equation, μ represents the unknown average sales per customer for the wholesaler's 2,000 customers.

Since the prior mean $M_0 = 27$ is greater than the break-even value $K = 25$, we know that the alternative "Buy the lot" is preferable. The expected profit is

$$\begin{aligned}E(\pi) &= -10,000 + 400M_0 = -10,000 + 400(27) \\ &= 800 \text{ dollars}\end{aligned}$$

Further, we can determine the expected value of perfect information, as we did in Chapter 10, page 235:

$$\text{EVPI} = tSL_N(D) \quad \text{where } D = \left| \frac{K - M}{S} \right| \quad (3)$$

Here M is the mean of the betting distribution; S is the standard deviation⁵; t is the slope of the loss function; and $L_N(D)$ is found in Appendix E. Using the prior mean, $M_0 = 27$, and standard deviation, $S_0 = 5$, we have

$$D = \left| \frac{25 - 27}{5} \right| = 0.4$$

$$L_N(D) = L_N(0.4) = 0.2304 \quad \text{from Appendix E}$$

and

$$\text{EVPI} = 400(5.0)(0.2304) = 461$$

That is, the prior expected value of perfect information is \$461.

Suppose that the wholesaler in question decided to obtain additional information in this decision problem by selecting a random sample of 50 customers (from the total of 2,000 customers) and asking each customer how many units he would purchase. Let us suppose that the average of these 50 "purchase orders" is 26.0 units per customer with a standard deviation of 14.14 units. Using symbols for sample data, $\bar{X} = 26.0$, $s = 14.14$, and $n = 50$ (sample size). The standard error of the sample mean can then be estimated as⁶

$$\begin{aligned} s_{\bar{X}} &= \frac{s}{\sqrt{n}} \\ &= \frac{14.14}{\sqrt{50}} = 2.0 \text{ units} \end{aligned}$$

Since the prior mean (M_0) and the sample mean (\bar{X}) are both above the break-even value ($K = 25$ units), there would be no reason

⁵ This notation differs from that used for EVPI in Chapter 10. In that chapter, μ and σ were the parameters of the normal betting distribution. In this chapter, these symbols describe the population to be sampled, and M and S (with subscripts 0 and 1) represent the parameters of the prior and posterior distributions.

⁶ Note that if the sample contains more than 5 percent of the population, the finite population correction factor should be included in estimating $s_{\bar{X}}$. That is, $s_{\bar{X}} = (s/\sqrt{n})(\sqrt{1 - n/N})$, where N is the population size.

to reverse the prior decision to buy the lot of merchandise. However, let us determine the posterior distribution anyway.

From Equation 1 we have

$$M_1 = \frac{\frac{M_0}{S_0^2} + \frac{\bar{X}}{\sigma_x^2}}{\frac{1}{S_0^2} + \frac{1}{\sigma_x^2}} = \frac{\frac{27}{5^2} + \frac{26}{2^2}}{\frac{1}{5^2} + \frac{1}{2^2}} = 26.14$$

From Equation 2,

$$\frac{1}{S_1^2} = \frac{1}{S_0^2} + \frac{1}{\sigma_x^2} = \frac{1}{5^2} + \frac{1}{2^2} = 0.29$$

Then

$$S_1^2 = 1/0.29 = 3.45 \text{ and}$$

$$S_1 = \sqrt{3.45} = 1.86$$

The values of $M_1 = 26.14$ and $S_1 = 1.86$ characterize the posterior betting distribution. After the sample, the decision-maker's best guess of the value of μ (mean sales per customer) is 26.14 units with a standard deviation of 1.86 units per customer. The posterior distribution is normal, indicating for example that the decision-maker should be willing to bet, with chances of 2 out of 3, that μ will be within the range 26.14 ± 1.86 or 24.28 to 28.00.

The posterior expected profit is

$$\begin{aligned} E(\pi) &= -10,000 + 400M_1 \\ &= 10,000 + 400(26.14) = \$456 \end{aligned}$$

And the posterior EVPI is determined as follows:

$$\begin{aligned} D &= \left| \frac{K - M_1}{S_1} \right| = \left| \frac{25.0 - 26.14}{1.86} \right| = 0.61 \\ L_N(D) &= 0.1659 \text{ from Appendix E} \\ \text{EVPI} &= tS_1L_N(D) = (400)(1.86)(0.1659) = \$123 \end{aligned}$$

Note that the posterior EVPI is considerably reduced from prior EVPI, even though the posterior mean M_1 was moved closer to the break-even point K . This resulted from the large reduction in standard deviation

from $S_0 = 5.0$ to $S_1 = 1.86$, so that there is considerably less chance for a large loss (i.e., for a value of μ considerably below $K = 25$).

It is important to recall that the posterior distribution in the example above was the result of a particular sample ($\bar{X} = 26$, $s = 14.14$, $n = 50$). A different sample result would have led to a different posterior distribution.

EVALUATION OF SAMPLING INFORMATION

In the above section we answered the following question: "Given that a sample has been taken, how should we use the information in the decision process?" We now turn to a different question: "Should we take a sample at all, and if so, how large should the sample be?" We shall answer the above question in two stages: first, we shall calculate the economic worth of a sample of a given size; second (in the next section), we shall determine the optimum sample size, which may be zero, so that no sample is warranted. Additional information, including sample evidence, has value to the decision-maker only if there is some chance that the information might change the prior decision. This implies that sample information generally enables us to reduce uncertainty (i.e., posterior expected loss).

Under the assumptions that we have been using in this chapter (two-action problem, linear profit functions, normal prior and sampling distributions), the evaluation of the economic worth of a sample can be accomplished in the six steps below, culminating in Equation 5.

Step 1: Determine the Prior Distribution. The decision-maker first finds the mean M_0 and standard deviation S_0 of his prior betting distribution.

Step 2: Determine the Profit Functions. The linear profit (or cost) functions are next determined. This includes the calculation of the break-even value K and the slope t of the opportunity loss functions.

Step 3: Estimate the Accuracy of the Proposed Sample. Accuracy is measured in terms of the sampling error ($\sigma_{\bar{x}}$) that we expect to obtain with the sample. Since the standard error $\sigma_{\bar{x}}$ is equal to σ/\sqrt{n} , we must have some estimate of σ , the standard deviation of the population from which the sample is to be taken.⁷ This estimate may be obtained from past studies of the population or similar populations, from a pilot sample taken to make such an estimate, or from an educated guess.

⁷The above formula for sampling error is for simple random sampling. More complicated formulas are necessary for different methods of sampling (e.g., stratification or cluster sampling); see Chapter 14.

Step 4: Estimate the Variance of the Posterior Distribution. This is determined from the prior variance S_0^2 (Step 1) and the sampling error estimate σ_x^2 (Step 3); that is, from Equation 2:

$$\frac{1}{S_1^2} = \frac{1}{S_0^2} + \frac{1}{\sigma_x^2}$$

Step 5: Determine the Variance Reduction. Designate a quantity S_*^2 which is obtained as follows:

$$S_*^2 = S_0^2 - S_1^2 \quad (4)$$

Note that S_*^2 is a measure of the reduction in the prior variance as a result of taking the sample. Thus, it is a measure of the value of the sample in reducing prior uncertainty.

Step 6: Calculate EVSI. The value of the sample in economic terms is given by the *expected value of sample information* or EVSI.⁸

$$\text{EVSI} = t S_* L_N(D) \quad \text{where } D = \left| \frac{K - M_0}{S_*} \right| \quad (5)$$

The symbol t represents the slope of the opportunity loss functions; M_0 is the prior mean; K is the break-even point; $L_N(D)$ is tabled in Appendix E; and S_* is obtained from Step 5 above.

The expected value of sample information is a measure of the expected additional profit that will be achieved by acting after the sample has been taken (and using the sample information) rather than acting before sampling. It is an expected value since different sample results will increase posterior profit by differing amounts or may even decrease expected posterior profit.

An Example

Let us continue the example of the wholesaler from page 381. Suppose that the wholesaler had not taken the sample discussed above but was considering the possibility of taking such a sample, say of 50 items, from his 2,000 customers. He would obtain advance orders from the 50 sample customers. Let us follow through the steps in obtaining EVSI in this illustration.

⁸ This formula is identical to that for EVPI with S_* replacing S_1 .

Step 1. Recall that the wholesaler had a normal prior distribution with mean $M_0 = 27$ items per customer and standard deviation $S_0 = 5$ items.

Step 2. The profit equations were

$$\begin{aligned}\text{Action "Buy the lot"} \quad \pi &= -10,000 + 400\mu \quad \text{in dollars} \\ \text{Action "Do not buy"} \quad \pi &= 0\end{aligned}$$

where μ is the unknown average sales per customer. We have previously determined the prior expected profit, $E(\pi) = \$800$, and the prior EVPI = \$461. The break-even value K is 25 items per customer, and the slope of the loss function $t = \$400$.

Step 3. We next need an estimate of σ , the standard deviation of potential orders from the population of 2,000 customers. Let us suppose that from past experience with similar items the wholesaler estimates σ at 14.14 units per customer. Then we can estimate the sampling error for a sample of size $n = 50$ as

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{14.14}{\sqrt{50}} = 2.0$$

Step 4. We then estimate the posterior variance as

$$S_1^2 = \frac{1}{\frac{1}{S_0^2} + \frac{1}{\sigma_x^2}} = \frac{1}{\left(\frac{1}{5.0^2}\right) + \left(\frac{1}{(2.0)^2}\right)} = 3.45$$

The posterior standard deviation is

$$S_1 = \sqrt{3.45} = 1.86$$

Step 5. The reduction in the prior variance due to sampling is

$$S_*^2 = S_0^2 - S_1^2 = (5.0)^2 - (1.86)^2 = 21.55$$

$$S_* = \sqrt{21.55} = 4.64$$

Step 6. The calculation of EVSI follows:

$$D = \left| \frac{K - M_0}{S_*} \right| = \left| \frac{25 - 27}{4.64} \right| = \left| \frac{2}{4.64} \right| = 0.431$$

$$L_N(D) = L_N(0.431) = 0.2200 \quad \text{from Appendix E}$$

$$\text{EVSI} = tS_*L_N(D) = (400)(4.64)(0.2200) = \$408$$

The value of the sample of 50 items to the decision-maker (the wholesaler in this example) is \$408. That is, we would expect a sample of this size to reduce uncertainty and to increase posterior expected profit by \$408. Recall that the expected value of perfect information is \$461. Thus, even such a moderate size sample gives close to perfect information (since \$408 is almost 90 percent of \$461).

Factors Influencing EVSI

The size of the expected value of sample information depends on some of the same factors that influence EVPI. In particular, both EVSI and EVPI vary directly with the slope of the loss function (t), the closeness of the prior mean to the break-even point ($|K - M_0|$), and the amount of uncertainty shown by the prior standard deviation (S_0). In addition, EVSI depends upon the sample size (n) and the dispersion in the sampled population (σ). The larger n , the larger EVSI; but the larger σ , the smaller EVSI since the sample will have relatively less precision.

OPTIMAL SAMPLE SIZE

In the previous section we assumed a fixed sample size and determined the economic worth of the sample. We now ask the question: "How large should the sample be, including the possibility of $n = 0$, no sample at all?" This is a matter of comparing the value of the sample (EVSI) with the cost of sampling.

Generally, the cost of sampling increases as a linear function of sample size as shown in Chart 16-1.

Chart 16-1

SAMPLING COSTS

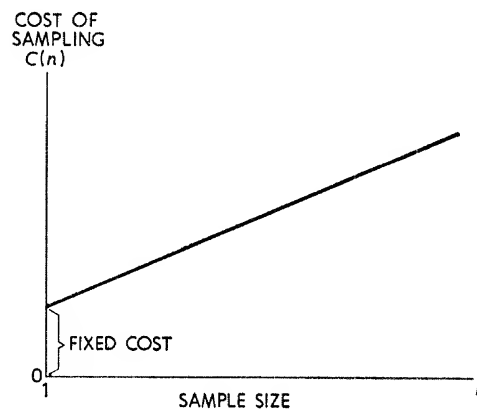


Table 16-1

CALCULATION OF EVSI FOR SELECTED VALUES OF n
(WHOLESALE'S DECISION TO BUY MERCHANDISE)

n	$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$	$S^2 = \frac{1}{\frac{1}{S_0^2} + \frac{1}{\sigma_{\bar{X}}^2}}$	$S_* = \sqrt{S_0^2 - S^2}$	$D = \left \frac{K - M_0}{S_*} \right $	EVSI = $t S_* L_N(D)$
20*	10.0	7.15	4.22	0.474	\$342
50	4.0	3.45	4.64	0.431	408
80	2.5	2.27	4.79	0.417	430
100	2.0	1.85	4.81	0.415	434
200	1.0	0.96	4.93	0.405	451

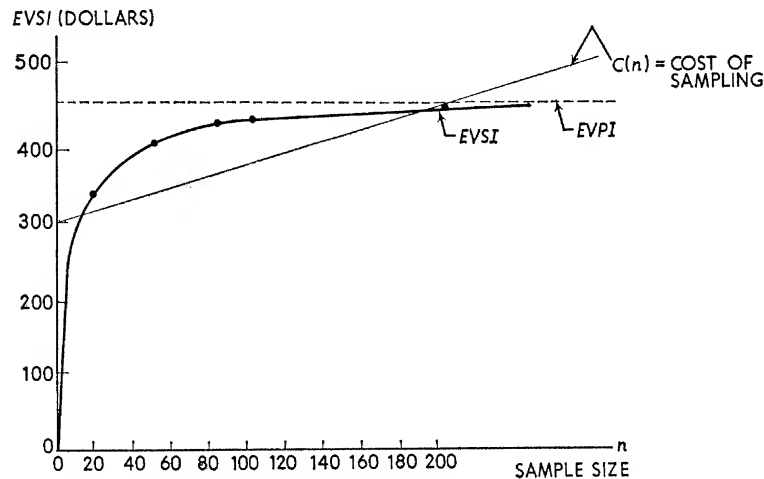
* Actually, for samples as small as $n = 20$, the sampling distribution of \bar{X} may not be normal when sampling from a skewed population. Hence, the calculation of EVSI, as shown in Table 16-1, is not, strictly speaking, accurate since the normality of the sampling distribution of \bar{X} is assumed.

The expected value of sample information is also a function of sample size. The larger the sample, the larger EVSI. In Table 16-1 the calculations for EVSI are shown for selected sample sizes for the example above (the wholesaler who is deciding about buying a lot of merchandise).

In Chart 16-2, EVSI is plotted as a function of the sample size n , with a smooth freehand curve drawn connecting the points calculated in Table 16-1, together with the point $n = 0$, for which $\text{EVSI} = 0$. Note that EVSI approaches the expected value of perfect information (EVPI) for very large values of n .

Chart 16-2

EXPECTED VALUE OF SAMPLE INFORMATION
AND COST OF SAMPLING
(WHOLESALE'S DECISION TO BUY MERCHANDISE)



Let us suppose that it would cost \$300 to set up the sample (a fixed cost) plus 75 cents per item included in the sample. Thus, the sampling cost can be expressed by the equation:

$$C(n) = \$300 + \$0.75 n$$

This equation is also shown in Chart 16-2. From this chart it can be seen that the value of the sample (EVSI) is greater than the cost for values of n between approximately $n = 15$ and $n = 200$. Hence, a sample with size somewhere between 15 and 200 would be preferable to no sample at all.

Let us define ENG S as the expected net gain from sampling, where

$$\text{ENG S} = \text{EVSI} - C(n) \quad (6)$$

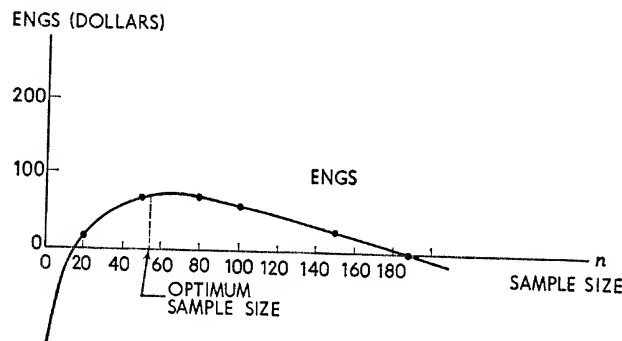
for any given value of n .

ENG S represents the difference between the economic worth of the sample information and the cost of obtaining the information. A small sample may not provide sufficient information to justify its cost. And since the additional value of sample information tends to decline as the sample size increases, a point is reached for large samples where, again, the sample value does not justify its cost. In between, sampling is worthwhile.

The ENG S for our example is plotted in Chart 16-3 as a function of the sample size n . ENG S is maximized at a value of about $n = 50$. This is the optimum sample size.⁹ The value of the sample exceeds the

Chart 16-3

EXPECTED NET GAIN FROM SAMPLING



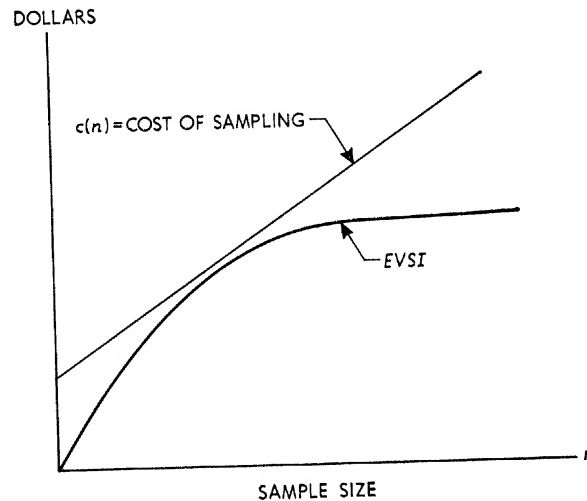
⁹ The determination of optimum sample size in situations such as the above can be more exactly determined. See R. Schlaifer, *Introduction to Statistics for Business Decisions*, Chap. 21. The calculated optimum for the above problem is $n = 52$.

sample cost by more at this point ($n = 50$) than at any other. Note that ENG_S is rather flat in the range $n = 40$ to $n = 80$, indicating that any sample size over this range would be only slightly less valuable than the optimum.

It may happen that $C(n)$ is greater than $EVSI$ for all values of n , as illustrated in Chart 16-4. Since the value obtained from sampling ($EVSI$) never exceeds the sampling cost, no sample should be taken.

Chart 16-4

EXPECTED VALUE OF SAMPLE INFORMATION
AND COST OF SAMPLING: SPECIAL CASE



The decision-maker should act with only his prior information (or find some less expensive means of obtaining information).

SUMMARY

Previous chapters developed the basic framework for combining probabilities, economic information, and sample results to determine optimal decisions. This chapter presents a special case of this general process, which has wide applicability.

There are four distributions involved in the analysis:

1. The *population from which the sample is to be drawn* can be of any type, and the mean of this distribution μ is unknown.
2. The *prior distribution* represents the decision-maker's judgment about the true mean μ of the population to be sampled.

3. The *sampling distribution* is the distribution of sample means \bar{X} about the true population mean μ . It represents the sampling error associated with estimating μ from the sample mean.
4. The *posterior distribution* represents the decision-maker's judgment about the true mean μ after the information of the sample has been incorporated.

The assumptions made in this chapter are

1. The prior distribution is normal.
2. The sampling distribution of \bar{X} is normal. This assumption will be satisfied if a large sample is taken.
3. The decision problem involves a choice between two acts, and the profits (or costs) may be expressed as a linear function of the unknown population mean μ .

If assumptions 1 and 2 above are satisfied, the posterior distribution is normal. And adding assumption 3 enables us to express the expected profit and the expected value of perfect information in simple formulas.

In order to determine if a sample should be taken, and how large it should be, we estimate the *expected value of sample information* (EVSI). This amount represents the expected economic worth of the sample in improving the decision about to be made. With the assumptions above, the calculation of EVSI for a given sample size n can be reduced to simple formulas.

To determine the *optimum sample size*, the cost of the sample must be balanced against its value. The *expected net gain from sampling* (ENGs) is the difference between EVSI and the sampling cost for a given size sample n . If ENGs is plotted on a chart for different values of n , the optimum sample size can be determined at the point where ENGs is largest. If ENGs is always negative, the cost of sampling exceeds its value for all n and no sample should be taken.

FORMULAS

Mean of the posterior distribution
for normal prior and normal
sampling distributions

$$M_1 = \frac{\frac{M_0}{S_0^2} + \frac{\bar{X}}{\sigma_{\bar{X}}^2}}{\frac{1}{S_0^2} + \frac{1}{\sigma_{\bar{X}}^2}}$$

$$\begin{array}{l} \text{Reciprocal of posterior} \\ \text{variance for normal} \\ \text{prior and sampling distributions} \end{array} \quad \frac{1}{S_1^2} = \frac{1}{S_0^2} + \frac{1}{\sigma_{\bar{x}}^2}$$

$$\begin{array}{l} \text{Expected value of} \\ \text{sample information} \end{array} \quad \text{EVSI} = tS_*L_N(D)$$

$$\begin{array}{l} \text{where } D = \frac{|K - M_0|}{S_*} \\ \text{and } S_*^2 = S_0^2 - S_1^2 \end{array}$$

$$\text{Expected net gain from sampling} \quad \text{ENGSI} = \text{EVSI} - C(n)$$

PROBLEMS

1. Discuss:
 - a) The meaning of a normal decision-making distribution.
 - b) Why sample information has value.
 - c) The distinction between a prior and a posterior distribution.
 - d) The effect of sample size on EVSI.
2. Determine the parameters of the posterior distribution in *a* through *d* below. Assume a normal prior with mean M_0 and standard deviation S_0 and a sample of size n with mean \bar{X} and standard deviation s .
 - a) $M_0 = 100, S_0 = 15; \quad \bar{X} = 90, s = 25, n = 100.$
 - b) $M_0 = 42, S_0 = 4; \quad \bar{X} = 43, s = 20, n = 35.$
 - c) $M_0 = 100, S_0 = 5; \quad \bar{X} = 90, s = 25, n = 30.$
 - d) $M_0 = 60, S_0 = 3; \quad \bar{X} = 55, s = 10, n = 100.$
3. A decision-maker has a prior normal distribution with mean $M_0 = 85$ and standard deviation $S_0 = 18$. The standard deviation of the population to be sampled is known to be 50. How large a sample must be taken so that the posterior standard deviation S_1 will be 4?
4. An election is about to be held in a large plant to see if the workers wish to be represented by a union. Management's expectations about the proportion of workers who will vote for the union is approximately normally distributed. Management feels that there is an equal chance that the proportion voting for unionization will be either above or below 40 percent. It also feels that there is an equal chance that the proportion voting for unionization will be within the range $33\frac{1}{3}$ to $46\frac{2}{3}$ percent as outside this range. A sample of 200 workers is selected at random and their voting intentions are determined. Ninety-six indicate that they will vote for unionization.
 - a) Describe the probability distribution that management should assign

to the event "Proportion of workers voting for unionization" after the sample has been taken. (Use normal approximation to binomial.)

- b) Based upon this probability distribution, what is the probability that the union will win the election?
 - c) What is the probability that the union will win the election, ignoring management's prior judgment and utilizing the sample information only?
5. An employer is concerned with hiring persons who are proficient in a certain manual skill measured on a scale between 0 and 100. The distribution of this skill among applicants for jobs is known to be normal with mean 50 and standard deviation 10. A test is used for screening purposes as a measure of this manual skill. However, the test is not perfect. The error associated with the test (difference between test score and "true" ability) is normally distributed with 0 mean and standard deviation 5 points.
- An applicant drawn at random scores 60 on the test.
- a) What is the probability that his "true" manual ability is below 50?
 - b) What is the probability that his "true" ability is above 60?
- (Hint: Treat the "true" distribution of skills as the prior distribution and the test error as the sampling distribution. The questions *a* and *b* apply, then, to the posterior distribution.)
6. Refer to the example in the text on pages 381 to 383. Suppose that a sample of 40 customers had been taken with a sample mean $\bar{X} = 24$ and standard deviation $s = 16$.
- a) Determine the posterior distribution.
 - b) What is the optimum action after the sample and what is the posterior expected profit?
 - c) What is the posterior EVPI?
7. Refer to the example in the text on pages 385 to 387.
- a) Calculate EVSI for a sample of 40 customers.
 - b) What is ENGSI for $n = 40$?
8. Refer to the example in the text on pages 385 to 387. This exercise is a study of the factors influencing EVSI. In each of *a* through *f* below, calculate EVSI for a sample size $n = 50$ with the indicated change and compare the result with that obtained in the text example. Add a sentence or two to explain the comparison.
- a) Suppose S_0 , the prior standard deviation, was 10 rather than 5.
 - b) Suppose S_0 , the prior standard deviation, was 3 rather than 5.
 - c) Suppose the prior mean M_0 was 25 rather than 27.
 - d) Suppose the prior mean M_0 was 32 rather than 27.
 - e) Suppose the standard deviation of the population to be sampled was $\sigma = 20$ rather than 14.14.
 - f) Suppose the standard deviation of the population to be sampled was $\sigma = 10$ rather than 14.14.

9. The Delta Company is considering the introduction of a new product. Delta distributes its products through 8,000 retail outlets. Management expressed its uncertainty about the demand for the new product in terms of a normal probability distribution with an unknown value of μ being the average sales in units per outlet. The mean of this prior distribution was 50 units per outlet, and the standard deviation was 15 units per outlet.

The new product would involve fixed costs of \$100,000 for machinery, promotion, advertising, and working capital. The incremental contribution (price less variable cost) from the sale of each unit was expected to be 22 cents.

- a) Using the above information, what is the best decision—to market or not market the new product? What is EVPI?
- b) Suppose management was considering taking a sample of 100 of the 8,000 retail outlets. The product would be introduced at each of the sampled outlets and the sales would be noted. The average sales per outlet in the sample would then be used as an estimate of the average sales for all 8,000 retail outlets. From past experience, the standard deviation of sales per outlet was estimated at 30 units. What is the EVSI for the sample of 100 outlets?
- c) Suppose the sample of 100 items was actually taken with the following result:

$$\begin{aligned}\bar{X} &= 59.2 \text{ unit sales per outlet} \\ s &= 28.7 \text{ unit sales per outlet}\end{aligned}$$

What action should be taken posterior to the sample? What is the posterior expected profit? What is the posterior EVPI?

10. Refer to Problem 9 above. Suppose a second sample of 50 outlets was being considered, after the first sample results in part c had been incorporated in the decision analysis. Should this second sample be taken if the cost of sampling is \$200, plus \$20 per outlet sampled?
11. As a dealer in retail hardware you are considering buying out the inventory of a merchant who is going out of business. You have a list of the items that he carried in stock but no exact inventory count has been made. There is the added problem of evaluating the worth of these items since many are obsolete or so old and damaged that they are valueless. Accordingly, you decide to take a sample of the items, check the count, and carefully value the sampled items.

Before taking the sample you examine the inventory. The owner is asking \$225,000 for the lot. You feel, on the basis of your cursory investigation, that it is worth \$235,000 to you, but there is much uncertainty about this guess. You feel that there is about 1 chance in 3 that your guess could be off as much as \$20,000 or more (either high or low).

There are 4,000 different items in the merchant's stock. You estimate that the standard deviation of value by item in the inventory is \$50.

Suppose further that the cost of taking a sample of any given size can be described by the equation:

Sampling cost = $\$150 + \$8n$ where n is the sample size

Ignore finite population correction factors throughout to simplify calculations. [*Hint:* Be sure to express your sampling unit and your inventory dollars in the same unit—e.g., since inventory is the total dollars, convert the sample estimate to total dollars (total = $N\bar{X}$) and for the error of the sample estimate ($s_{\text{total}} = Ns_{\bar{X}}$).]

- a) Before consideration of sampling, would you buy the merchandise? What is EVPI? (Assume normality.)
- b) What size sample (if any) should be taken? Explain.

12. The Ivanhoe Construction Company has been offered a contract to build a plant for the Zeta Steel and Wire Company. A contract price of \$2.8 million has been agreed upon by both parties. Mr. Ivanhoe, the president, has estimated that his cost will be \$2.4 million, leaving a profit before taxes of \$400,000.

However, Zeta is fearful of losing ground to its competitors and is in a considerable hurry for its new plant. Zeta proposes an incentive contract that would reward Ivanhoe with \$50,000 for each month that the project was completed before the scheduled date (20 months from now) and a penalty of \$50,000 for each month beyond the target date.

Mr. Ivanhoe is somewhat dubious about agreeing to this provision in the contract. He feels that the contract can be completed in the agreed time (20 months), or even shorter, if all goes well, but unexpected shortages of materials or other contingencies could considerably delay the project. When questioned further, Mr. Ivanhoe said that 21 months was his "best guess" as to completion time. This would allow for some unplanned delays. He further felt that chances were good (say 2 chances out of 3) that the completion date would not vary more than 3 months either way from his guess.

Ivanhoe had an alternative venture that would give a before-tax profit of \$300,000. This alternative would have to be foregone if the Zeta project were undertaken.

- a) Assume a normal distribution for the time to complete the Zeta project. Based upon this, what should Ivanhoe do and what is the expected profit?
- b) Do you think that the assumption of normality is reasonable in this case? Why or why not? If the distribution were not normal, how would it affect your answer to part a above?

Mr. Ivanhoe had been studying the possibility of using some "critical path" technique (such as PERT or CPM) as an aid in controlling and predicting schedules. Ivanhoe contracted Mr. Wade of a local consulting firm specializing in critical path methods. After examining Ivanhoe's problem Wade indicated that, using his methods, he could make a reasonably accurate estimate of the time to complete the construction project. This estimate would not be perfectly accurate since all contingencies could not be planned for. Based upon his experience with similar projects, Wade felt that he could estimate completion time within ± 1 month with 80 percent probability. Wade's consulting fee for this estimate would be \$40,000.

- c) Should Mr. Ivanhoe hire Mr. Wade to make an estimate of time to complete the project before Ivanhoe decides to accept or reject the Zeta Steel and Wire contract?

SELECTED READINGS

BIERMAN, H.; BONINI, C.; FOURAKER, L.; and JAEDICKE, R. *Quantitative Analysis for Business Decisions*. Homewood, Illinois: Richard D. Irwin, 1965.

Chapters 8 and 9 deal with revising probabilities and decision-making for normal distribution. Chapters 12 and 13 deal with revising binomial probabilities.

CLELLAND, RICHARD C.; DE CANI, JOHN S.; BROWN, FRANCIS E.; BURSK, J. PARKER; and MURRAY, DONALD S. *Basic Statistics with Business Applications*. New York: John Wiley, 1966.

Chapter 16 treats Bayes' Theorem, including the revision of probabilities in the two-action, linear-cost situation.

HADLEY, G. *Introduction to Probability and Statistical Decision Theory*. San Francisco: Holden-Day, 1967.

Chapter 9 is an extended treatment of Bayes' Theorem and its application to decision problems, with a somewhat different approach.

LINDLEY, D. V. *Introduction to Probability and Statistics from the Bayesian Viewpoint*. 2 vols. Cambridge, England: The University Press, 1965.

A mathematical treatment, with emphasis on probability distributions and inferences for normal distributions.

PRATT, JOHN W.; RAIFFA, HOWARD; and SCHLAIFER, ROBERT. *Introduction to Statistical Decision Theory*. Preliminary edition. New York: McGraw-Hill, 1965.

An advanced treatment of Bayesian decision theory.

SCHLAIFER, R. *Introduction to Statistics for Business Decisions*. New York: McGraw-Hill, 1961.

Part 2 deals with binomial sampling and Bayes' Theorem. Part 4 deals with sampling, revision of normal probabilities, and decision-making with normal distributions.

17. PROBABILITY MODELS AND DECISION-MAKING

IN CHAPTERS 9, 10, 15, and 16 probability distributions have been used to represent uncertainty about unknown variables. We then adopted a *general* approach to decision-making under uncertainty. In this chapter we shall consider some special decision situations for which *specific* probability models have been developed. Our purpose is to study the process of building probability models that are useful in making business decisions.

We shall not go into each class of decision model in depth, for this would take several volumes. Rather, this chapter is a brief survey intended to demonstrate the broad usefulness of some of the many probability models that have been developed.

A BIDDING MODEL

Consider the plight of a contractor who must submit a bid on a contract in competition with several other bidders. The contract is to be awarded to the lowest bidder. Suppose the contractor has made an estimate of his cost to do the work involved. This would represent his lowest bid.¹ The higher the contractor raises his bid, the more his profit, but the less his chances of winning. The contractor must find some balance between profit on the contract and the probability of winning.

As an example, suppose Contractor Jones is bidding on a job that he expects would cost him \$500,000 to complete. Jones has excess capacity

¹One can imagine situations in which a contractor might bid below his cost estimate. If he has excess capacity, he could use marginal costs as the decision amount. Further, one might bid low on a research and development contract, e.g., with the expectation of obtaining a profitable procurement contract later.

Table 17-1

PROBABILITIES OF VARIOUS BIDS

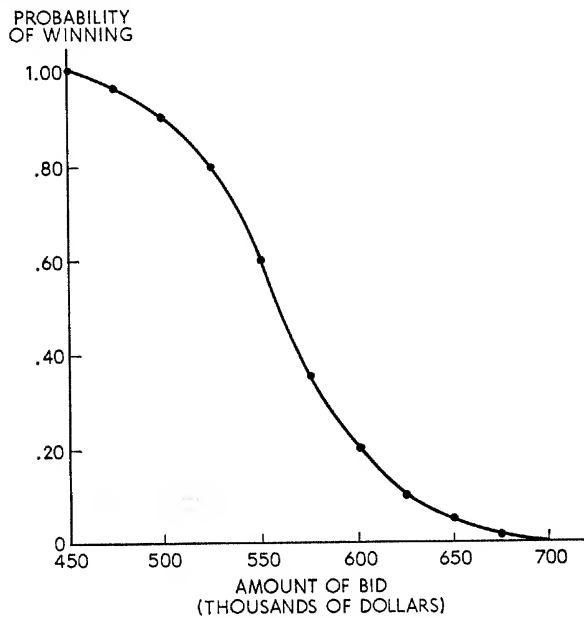
Jones' Possible Bid	Jones' Subjective Probability that His Bid is Lowest	Cumulative Probability of Winning with Bid
\$450,000	0.05	1.00
475,000	0.05	0.95
500,000	0.10	0.90
525,000	0.20	0.80
550,000	0.25	0.60
575,000	0.15	0.35
600,000	0.10	0.20
625,000	0.05	0.10
650,000	0.03	0.05
675,000	0.02	0.02
Total.....	1.00	

and can take on the new job. Several other contractors are also bidding on the job. Jones has bid against these contractors for jobs in the past, and he assigns the probabilities shown in Table 17-1 about the lowest bid of his competitors.

As can be seen from Table 17-1, Jones has estimated the subjective

Chart 17-1

CUMULATIVE PROBABILITY OF WINNING WITH BID INDICATED



probabilities of winning for any bid amount. These data are plotted in Chart 17-1, and a smooth curve is drawn to connect the points. This gives the probabilities of winning for bids intermediate to those shown in Table 17-1.

Jones can then determine his expected profit for any bid by multiplying the profit for each bid (if it wins) by the probability of winning with this bid. Then he should select the bid with the highest expected profit, according to our decision criterion developed in Chapter 9.

In Table 17-2, as the contractor increases the bid, the expected profit

Table 17-2
EVALUATION OF EXPECTED PROFIT

Bid	Profit if Bid Wins	Probability of Winning	Expected Profit
\$500,000	0	0.90	0
525,000	\$25,000	0.80	\$20,000
550,000	50,000	0.60	30,000
575,000	75,000	0.35	26,250
545,000*	45,000	0.64*	28,800
555,000*	55,000	0.54*	29,700

* Interpolated values on either side of \$550,000, with probabilities from Chart 17-1.

goes up to \$30,000 at a bid of \$550,000 and then begins to decline. The table only shows the expected profits near the peak amount. The last two rows in Table 17-2 show additional values around \$550,000 to determine a more exact optimum, but in this case they merely confirm that the bid of \$550,000 is the best.

The most difficult part of the analysis in this bidding model is to estimate the probabilities of winning. Some information about this distribution can be obtained from past bidding situations, but in the final analysis the distribution rests upon the subjective judgment of the decision-maker.

AN INVENTORY MODEL

Consider a merchant who must decide how many units of a perishable product to purchase. Suppose he buys this product for c dollars per unit in the morning and then sells it during the day for p dollars. Any stock remaining unsold at the end of the day has no value and is thrown away. The decision problem is to select q , the optimum number of units to purchase.²

² For obvious reasons, this problem is referred to as the "newsboy" problem and the model suggested below as the "newsboy" model.

Let us suppose that the demand for the product on a given day is a random variable X with probability distribution $P(X)$. The merchant does not know exactly how many he can sell but knows the probabilities $P(X)$ for all values of X .

Of course, we can solve this problem by constructing a payoff table and proceeding as in Chapter 9. Assuming specific numerical values:

c = purchase cost = \$4 per unit

p = sales price = \$6 per unit

q = units stocked

X = demand—between 0 and 4 units

$P(X)$ = probability distribution of demand—see Table 17-3

Table 17-3

PAYOFF TABLE FOR INVENTORY PROBLEM

(DOLLARS PROFIT)

Event: Demand X	Probability $P(X)$	Actions: q				
		0	1	2	3	4
0	0.10	0	-4	-8	-12	-16
1	0.10	0	2	-2	-6	-10
2	0.20	0	2	4	0	-4
3	0.40	0	2	4	6	2
4	0.20	0	2	4	6	8
Total . . .	1.00					
Expected Profits	0		1.4	2.2	1.8	-1.0

The payoff table shows the profit for each combination of action and event. Thus, if the merchant buys two units and sells one, his "profit" is $6 - 8 = -\$2$. Multiplying these profits by the probabilities and adding the products, we get the expected profit for each action in the bottom row. The maximum expected profit of \$2.2 indicates that the optimal action is to purchase 2 items, that is, $q_* = 2$, where the $*$ indicates the optimum value of q .

The use of a payoff table, however, would be extremely cumbersome if the number of possible values of q was large. Fortunately, we can restructure the payoff table to achieve an easier solution.

Let us first look at the opportunity losses in this inventory situation. The best decision would be to purchase exactly the amount sold. A loss occurs either when more are purchased than demanded (this is the loss from overstocking, designated l_o) or when demand exceeds the number purchased (the lost profit is termed the loss from understocking, l_u). In

the above example, the loss from overstocking (l_o) is \$4 per unit, that is, the purchase cost of an unsold item. The loss from understocking is the profit that could have been made with the additional sale ($l_u = \$2$ per unit). *Note that the loss of overstocking or understocking is a constant amount per unit.* Thus, if two units are overstocked the loss will be $2l_o = 2 \times 4 = 8$. The fact that costs are linear makes a simple analytic solution possible.

An Analytic Solution: Discrete Functions

Let us consider the cumulative probability distribution $P(X < q)$, the probability that demand (X) will be less than q units. For the discrete probability distribution in Table 17-3, the cumulative distribution is shown in Table 17-4.

Table 17-4
CUMULATIVE PROBABILITY DISTRIBUTION
THAT DEMAND (X) WILL BE LESS THAN THE NUMBER STOCKED (q)

X or q	$P(X)$	$P(X < q)$
0	0.10	0
1	0.10	0.10
2	0.20	0.20
3	0.40	0.40
4	0.20	0.80
5	0	1.00

The optimal stock level is then obtained by first finding the *largest* value of q that satisfies the following equation:³

$$P(X < q) < \frac{l_u}{l_u + l_o} \quad (1)$$

That is, we search the third column of Table 17-4 until we find a value just less than the ratio $l_u/(l_u + l_o)$. The stock level associated with this value is the optimal stock.

In the example of the merchant above, $l_u = 2$ and $l_o = 4$ so that

$$\frac{l_u}{l_u + l_o} = \frac{2}{2 + 4} = \frac{1}{3}$$

In Table 17-4, column 3, the largest value of $P(X < q)$ that is less than one third is 0.20, corresponding to $q = 2$. Hence, the optimum

³ This relationship is given without proof. When $P(X < q)$ exactly equals $l_u/(l_u + l_o)$ for a given value of q , the optimal stock level can be either q or $(q - 1)$. Both alternatives have the same profit.

stock level is $q_* = 2$, which is the same as we obtained by the longer payoff table method above.

Continuous Functions

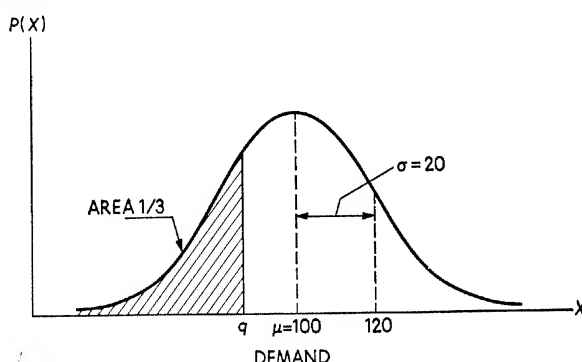
When $P(X)$, the probability distribution for demand, is represented as a continuous distribution, the optimum can be determined by finding the value of q such that

$$P(X < q) = \frac{l_u}{l_u + l_o} \quad (2)$$

This is the same formula as for discrete data, except for the equality sign.

Chart 17-2

NORMAL DEMAND DISTRIBUTION FOR INVENTORY PROBLEM



As an example, consider the situation in which $l_u = 2$ and $l_o = 4$, as above. Let demand be represented by a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 20$ (see Chart 17-2).

We seek a value of q such that $P(X < q) = l_u / (l_u + l_o) = 2 / (2 + 4) = 1/3$. This is equivalent to finding a value of q such that the area in the tail of the normal curve to the left of q is $1/3$. From the table of areas under the normal curve (Appendix D), we can determine that one third of the area lies to the left of a point 0.43 standard deviation below the mean (i.e., $\mu - 0.43\sigma$).⁴ Since our mean is 100 and $\sigma = 20$, the optimum value is

$$q_* = 100 - (0.43)20 = 91.4 \text{ or } 91 \text{ units}$$

⁴ Whether the value of q is above or below the mean depends upon whether $l_u / (l_u + l_o)$ is greater or less than 0.50.

Goodwill Costs and Scrap Allowances

The above analysis implicitly assumed that the only loss associated with understocking was the lost profit. In addition, when a stock shortage (an unfilled demand) occurs, there may be some loss of customer goodwill that will affect future profits. It is possible to include a quantitative measurement of this goodwill loss simply by adding some amount to the loss of understocking so that $L_u = \text{lost profit per unit} + \text{goodwill loss per unit}$. The new value for L_u can then be used in the model exactly as before.

The analysis presented above also assumed that the inventory had no value at the end of the period. This need not be the case. If, for example, the merchandise can be sold at a discounted or scrap price (e.g., day-old bread) at a later period, then the cost of overstocking is the purchase price less the discounted or scrap price. The new value of l_o (purchase cost less salvage value) can be inserted in Equations 1 and 2, and the procedures for determining q_* can be followed as above.

The above "inventory" model has wide applicability to many problems that do not actually involve inventories. The number of employees needed to handle a varying amount of work is an example of such a situation. The loss associated with understocking is the overtime premium that must be paid if too few employees are hired. The loss associated with overstocking is the pay of the idle workers when no work is available. The critical factor in the general application of this model is that the opportunity losses of overstocking and understocking must be linear—that is, a constant amount per unit for all units. Furthermore, this model is only one of a great many designed to represent inventory situations.

A QUEUING MODEL

Queues, or waiting lines, are common occurrences in many situations where there are random or unscheduled events. Waiting lines are familiar phenomena in barber shops, supermarkets, tool cribs in factories, telephone switchboards, repair shops, and a host of other situations. In all these cases, people, telephone calls, or machines "arrive" in a somewhat random fashion at a "service station" where they await their turn to be "serviced." The time taken to wait on or service an individual may also be a random variable. Queuing theory is the study of the probabilities associated with the length of the waiting line and the time an individual must wait in the queuing system.

There are several characteristics of queuing problems:

1. The pattern or probability distribution associated with the arrivals at the service center.
2. The probability distribution associated with the time taken to wait on or service an individual.
3. The queue discipline. The queue may be organized on a first-come—first-serve basis, on a random basis, or according to some priority scheme. Also, an individual may balk at entering the queue if it is too long.
4. The number of service channels. There may be only a single channel (e.g., one switchboard operator) or multiple channels (e.g., the several checkout counters in a grocery store).

With certain assumptions about these four factors, it is possible to analyze, in mathematical fashion, the behavior of the queue. For other sets of assumptions, mathematical results are not available and we must resort to simulation (see the last section of this chapter) for our analysis.

One Channel Model

Let us assume that arrivals occur in a random pattern and that the probability of an arrival in any unit of time is constant and is independent of the number of arrivals in previous periods. In Chapter 8 we saw that these were the assumptions of the Poisson process and, hence, the arrivals may be described by a Poisson distribution. The average number of arrivals per unit of time is m , the sole parameter of this distribution.

Let us assume that the varying number of customers serviced per unit of time also follows the Poisson distribution, with the same assumptions as above. The mean of this distribution is the service rate a . The *average service time*, the time taken on the average to service a customer, is $1/a$, the reciprocal of the service rate.

We will further assume that all arrivals will gain the queue (i.e., none will balk and go elsewhere). Hence, the average service rate a must be greater than the average arrival rate m ; otherwise, the queue will grow indefinitely large since individuals would be arriving faster than they can be serviced. Further, arrivals will be serviced on a first-come—first-serve basis.

We wish to study the behavior of this probabilistic system over time. Since no one is waiting when the system is opened, the queue starts out

at zero. The queue will expand and contract in a random pattern as time goes on. Soon, the effect of starting at scratch (no one in the queue) wears off and the queuing system reaches *equilibrium*. In equilibrium, the system responds only to the random pattern of arrivals and departures.

While the system in equilibrium behaves randomly in the sense that we cannot predict the *exact* queue length at any point in time, nevertheless, the system has certain predictable properties.⁵ In particular, we can find the *probability* for a queue of any length. The probability of exactly n individuals in the system (n = number in the queue waiting for service plus the one being serviced) is given by

$$P_n = \left(\frac{m}{\alpha}\right)^n P_0 \quad \alpha > m \quad (3)$$

where

$$P_0 = 1 - \frac{m}{\alpha} = \text{probability of no one in the system} \quad (4)$$

P_0 represents the probability that an individual will not have to wait for service when he arrives. P_n is the probability that he will find exactly n individuals ahead of him. From the probabilities shown in Equations 3 and 4, it is possible to determine certain other measures. The average or expected number of individuals in the system (n) is

$$E(n) = \frac{m}{\alpha - m} \quad (5)$$

The average or expected number of individuals in the queue or waiting line (n') excluding the individual being serviced is

$$E(n') = \frac{m^2}{\alpha(\alpha - m)} = E(n) \left(\frac{m}{\alpha}\right) \quad (6)$$

Let w be the time an individual spends in the system (i.e., in waiting plus being serviced). Then the average time that an individual will spend in the system is

$$E(w) = \frac{1}{\alpha - m} \text{ in units of the time interval selected} \quad (7)$$

⁵ The derivation of the equations for the queue behavior is not shown. See the references at the end of this chapter.

And the average or expected time that an individual will spend waiting in line (w') is:

$$E(w') = \frac{m}{\alpha(\alpha - m)} = E(w) \left(\frac{m}{\alpha} \right) \text{ in units of the time interval selected} \quad (8)$$

Equations 5 through 8 can prove helpful in the economic analysis of queuing situations, as will be demonstrated in the following example.

Example

An airline office has one reservation clerk to handle telephone calls for information and reservations. During the peak hours of the day (10 AM to 4 PM), calls arrive at random at an average rate of 10 per hour ($m = 10$). The clerk can handle 15 calls per hour on the average ($\alpha = 15$). The calls arriving and the clerk's completion of calls both follow Poisson distributions. If we assume that calls are answered on a first-come-first-serve basis, and that callers wait until the clerk is free, we can use the probability model described by Equations 3 through 8 above.

The probability that a caller will get immediate service is

$$P_o = 1 - \frac{m}{\alpha} = 1 - \frac{10}{15} = \frac{1}{3}$$

The average number of calls either waiting or being answered by the clerk is

$$E(n) = \frac{m}{\alpha - m} = \frac{10}{15 - 10} = 2$$

The average time that a customer must wait before being served is

$$E(w') = \frac{m}{\alpha(\alpha - m)} = \frac{10}{(15)(5)} = \frac{2}{15} \text{ hour or 8 minutes}$$

And the average time in which a caller could expect to complete his call (including both waiting time and service time) is

$$E(w) = \frac{1}{\alpha - m} = \frac{1}{15 - 10} = \frac{1}{5} \text{ hour or 12 minutes}$$

Suppose that the management of the airline is considering installing new equipment which would enable the reservation clerk to service 20

calls per hour ($\alpha = 20$) rather than the previous rate of 15 calls per hour. Let us investigate the effects of this change upon customer service. Note first that the probability of immediate service is increased:

$$P_o = 1 - \frac{m}{\alpha} = 1 - \frac{10}{20} = \frac{1}{2}$$

The average number of calls in the system is

$$E(n) = \frac{m}{\alpha - m} = \frac{10}{20 - 10} = 1$$

And the average waiting time before being served is reduced to

$$E(w') = \frac{m}{\alpha(\alpha - m)} = \frac{10}{(20)(10)} = \frac{1}{20} \text{ hour or 3 minutes}$$

In order to determine if the new system should be installed, the management of the airline would compare the reduction of 5 minutes in waiting time (from 8 to 3 minutes) with the cost of the new system. The saving of 5 minutes per call times 10 calls per hour times the 6 peak hours of the day gives a total reduction of 300 minutes in customer waiting time per day.

Suppose the new system would cost \$60 per day. Then, if the management attached a cost of 20 cents to each minute of customer waiting time, the new system would exactly break even (\$60 = 300 minutes \times 20 cents per minute). If management valued customer waiting time at more than 20 cents per minute, the new system should be installed.

Another alternative open to the management is to have a second reservation clerk so that two calls could be handled simultaneously. This would be a *two-channel* system. A simulation approach to analyzing such a situation is described in the next section. Mathematical methods of handling certain multichannel cases are described in the references at the end of the chapter.

SIMULATION

In the probability models described above it was possible to obtain an optimum solution by direct analysis. In many situations the models we build are of such complexity that we are not able to solve them by mathematical means. One method of analysis in such situations is to

build a simulated model and to study this model under various conditions. The engineers who build scale models of airplanes and test these models in wind tunnels are following this procedure. Similarly, the engineers who build replicas of dams on a small scale before beginning the large-size project are using the tool of simulation. And there are many more instances of how physical models are used to approximate real-world behavior.

It is also possible to build simulation models of many business processes. The procedure does not involve the construction of a physical model (such as a dam or model airfoil) but utilizes a symbolic or logical structure showing the relationships or connections between the important variables in the business situation. It is easier to understand the meaning of this if we use a specific example.

Simulation of a Queuing Situation

Consider a queuing situation in which the service time is constant. Suppose we know the distribution of arrivals, and we wish to compare the properties of this system for one- and two-channel operations. The mathematical queuing model presented earlier does not apply because of the constant service time in this example.⁶ Also, the mathematical model was limited to a one-channel case.

In particular, let the arrivals represent passengers checking in at an airport desk preparatory to departure. The arrival times of 50 passengers during a typical late afternoon rush period are listed in Table 17-5, columns 1 and 2 (time zero is 4 PM). With new communications equipment, management estimates that service time will be a constant 3 minutes per customer. The decision to be made is whether to provide for one or two clerks, or "channels." Let us investigate the effects of this sequence of arrivals on a one-channel system and on a two-channel system.

This is shown first for the one-channel case in the schematic diagram, Chart 17-3. Time is plotted along a continuous scale running down the length of the diagram. Arrivals are shown at the time they enter the system. They either go directly into service with no wait (for example, arrivals Nos. 1 and 3) or they must wait in the queue until the service channel is free. Arrival No. 2, for example, comes into the system at time 0:04.4. But service started on No. 1 at 0:04.3 and continues until 0:07.3, a three-minute service time. Thus, the service channel becomes

⁶ A mathematical analysis for Poisson arrivals, constant service times, and one-channel operation is described in R. Schlaifer, *Probability and Statistics for Business Decisions* (New York: McGraw-Hill, 1959), Chap. 19.

free at 0:07.3 and No. 2 can be serviced. The waiting time for No. 2 is thus 2.9 minutes (his starting service time 0:07.3 minus his arrival time 0:04.4). Note that an arrival may find more than one individual ahead of him. For example, No. 11 finds three individuals ahead of him (plus the one being serviced) when he arrives at time 0:30.0.

Since it is time-consuming to continue the schematic procedure employed in Chart 17-3, let us do the same thing in another form, Table 17-5. In this table the "Time Begun Service," column 3, for the one-channel case is simply either (1) the time of arrival or (2) the "Time Begun Service" for the previous arrival plus three minutes, *whichever is later*. This implies that an arrival can go directly into service if the channel is free or must wait until the immediately previous

Chart 17-3

SCHEMATIC DIAGRAM OF THE ONE-CHANNEL QUEUING SITUATION

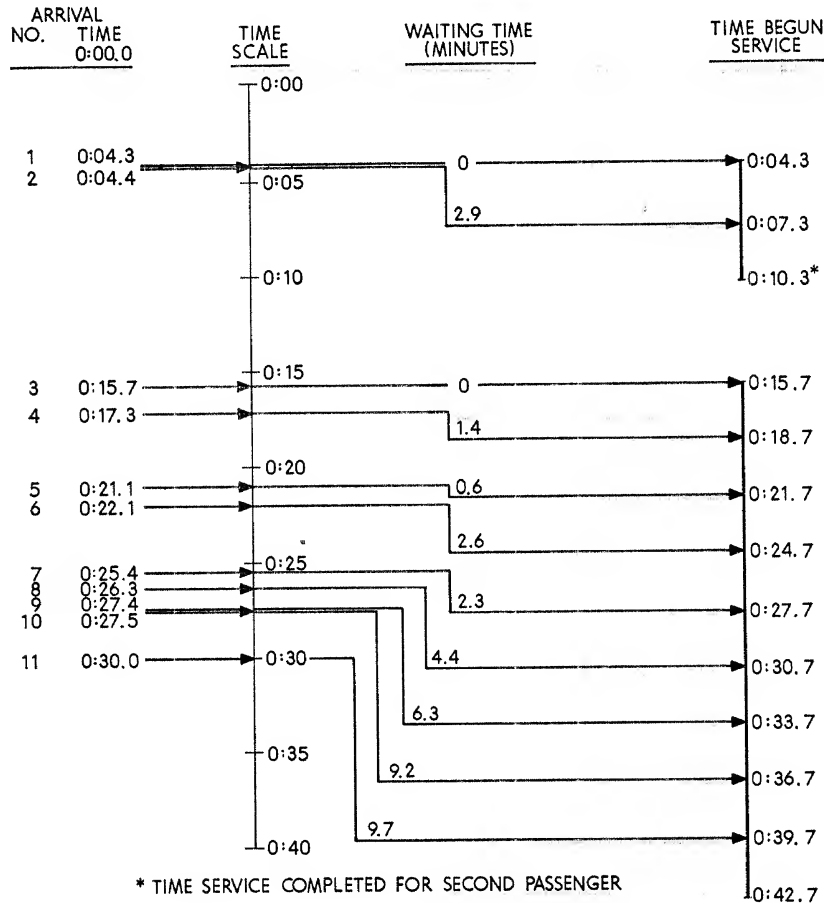


Table 17-5
SIMULATION OF QUEUING SITUATION

Arrival		One-Channel Case		Two-Channel Case	
(1) Arrival Number	(2) Time of Arrival	(3) Time Begun Service	(4) Waiting Time	(5) Time Begun Service	(6) Waiting Time
1	0:04.3	0:04.3	0	0:04.3	0
2	0:04.4	0:07.3	2.9	0:04.4	0
3	0:15.7	0:15.7	0	0:15.7	0
4	0:17.3	0:18.7	1.4	0:17.3	0
5	0:21.1	0:21.7	0.6	0:21.1	0
6	0:22.1	0:24.7	2.6	0:22.1	0
7	0:25.4	0:27.7	2.3	0:25.4	0
8	0:26.3	0:30.7	4.4	0:26.3	0
9	0:27.4	0:33.7	6.3	0:28.4	1.0
10	0:27.5	0:36.7	9.2	0:29.3	1.8
11	0:30.0	0:39.7	9.7	0:31.4	1.4
12	0:35.5	0:42.7	7.2	0:35.5	0
13	0:40.2	0:45.7	5.5	0:40.2	0
14	0:48.2	0:48.7	0.5	0:48.2	0
15	0:48.4	0:51.7	3.3	0:48.4	0
16	0:48.5	0:54.7	6.2	0:51.2	2.7
17	0:49.0	0:57.7	8.7	0:51.4	2.4
18	0:49.1	1:00.7	11.6	0:54.2	5.1
19	0:49.6	1:03.7	14.1	0:54.4	4.8
20	0:50.1	1:06.7	16.6	0:57.2	7.1
21	0:53.6	1:09.7	16.1	0:57.4	3.8
22	1:00.5	1:12.7	12.2	1:00.5	0
23	1:04.0	1:15.7	11.7	1:04.0	0
24	1:06.7	1:18.7	12.0	1:06.7	0
25	1:07.0	1:21.7	14.7	1:07.0	0
26	1:12.0	1:24.7	12.7	1:12.0	0
27	1:12.1	1:27.7	15.6	1:12.1	0
28	1:16.8	1:30.7	13.9	1:16.8	0
29	1:18.0	1:33.7	15.7	1:18.0	0
30	1:24.7	1:36.7	12.0	1:24.7	0
31	1:25.7	1:39.7	14.0	1:25.7	0
32	1:28.2	1:42.7	14.5	1:28.2	0
33	1:31.8	1:45.7	13.9	1:31.8	0
34	1:31.9	1:48.7	16.8	1:31.9	0
35	1:35.4	1:51.7	16.3	1:34.8	0.6
36	1:36.0	1:54.7	18.7	1:36.0	0
37	1:36.1	1:57.7	21.6	1:37.8	1.7
38	1:51.2	2:00.7	9.5	1:51.2	0
39	1:53.1	2:03.7	10.6	1:53.1	0
40	2:05.2	2:06.7	1.5	2:05.2	0
41	2:11.3	2:11.3	0	2:11.3	0
42	2:12.5	2:14.3	1.8	2:12.5	0
43	2:21.5	2:21.5	0	2:21.5	0
44	2:21.9	2:24.5	2.6	2:21.9	0
45	2:26.9	2:27.5	0.6	2:26.9	0
46	2:36.0	2:36.0	0	2:36.0	0
47	2:38.0	2:39.0	1.0	2:38.0	0
48	2:44.2	2:44.2	0	2:44.2	0
49	2:44.7	2:47.2	2.5	2:44.7	0
50	2:45.5	2:50.2	4.7	2:45.5	0
Sum of last 40 items.....		370.6		29.6	
Average wait.....		9.62		0.74	

arrival is finished with his service. The waiting time (column 4) is the difference between arrival time and the "Time Begun Service."

For the two-channel case, we use the same history of arrivals. However, the "Time Begun Service" (column 5) for, say, the n th arrival is now determined as (1) the time of arrival or (2) the "Time Begun Service" for the $(n - 2)$ th arrival (i.e., the arrival before last) plus three minutes, *whichever is later*.

Because there are two channels, an arrival will have to wait only if both channels are being utilized. And if both channels are in use, he must wait until the second arrival before him is finished before he can begin being serviced.

The waiting time (column 6) for the two-channel case is, as before, the difference between the arrival time and the "Time Begun Service" for each arrival.

In Table 17-5, we simulated the waiting times for 50 arrivals covering a period of about 165 minutes. Of course, we could continue the simulation for any number of arrivals. We wish to compare the performance of the one-channel system with the two-channel. We should like to make this comparison when both systems are in equilibrium, that is, when they have been operating long enough to be independent of initial conditions (e.g., starting the queuing process with no waiting line). For this reason we shall exclude the first 10 arrivals from our consideration. Comparing, then, the performance of the two systems for arrivals 11 through 50 we see that the average wait of 9.62 minutes with the one-channel system is reduced to 0.74 minute for the two-channel system. Of course, these estimates are based upon a relatively small sample of arrivals and we should carry out Table 17-5 for many more observations before making a decision about the relative merits of the one- versus two-channel systems.

Note that simulation, in this example, meant the portrayal on paper of a real-world system. The simulation model, as well as other models, can only approximate the elements of the real world, but where actual experience is difficult or impossible to obtain (e.g., why build a second channel to find if one is necessary?), a set of models involving different assumptions can provide an invaluable series of "dry runs."

The Monte Carlo Method of Simulating Probability Distributions

In the above example, the model utilized a record of actual arrivals of airline passengers. If no records are available, however (as in instituting a new process), we can still generate, in an artificial fashion, a time series or history that would have properties similar to those of a real-

Table 17-6

PROBABILITY DISTRIBUTION OF SALES

Daily Sales, Units	Probability	Cumulative Probability	Random Number Assignments
50	0.025	0.025	000 to 024
51	0.225	0.250	025 to 249
52	0.350	0.600	250 to 599
53	0.250	0.850	600 to 849
54	0.125	0.975	850 to 974
55	0.025	1.000	975 to 999
	1.000		

world series. If we know the probability distribution involved it is possible to generate such a series by a process known as *Monte Carlo* analysis.

As an example, suppose that the probability distribution for daily sales of a certain product is as shown in Table 17-6. Cumulative probabilities are listed in column 3.

Let us now assign three-digit numbers to each sales level in accordance with the cumulative probabilities. Thus, we assign the numbers from 000 through 024 (a total of 25 three-digit numbers) to the sales level of 50 units, and so on. We then proceed to draw three-digit random numbers from a table of random numbers. Each random number will determine a daily sales amount since each three-digit number is assigned to a sales level. The first random number drawn is 504. This falls in the group 250 to 599 that corresponds to sales of 52 units (see Table 17-6). The second random number is 113, which is in the group 025 to 249 and corresponds to sales of 51 units. We continue on with this process of drawing random numbers and generating a history of sales, as shown in Table 17-7.

Table 17-7

MONTE CARLO SIMULATION OF DAILY SALES

Day	Random Number	Sales
1	504	52
2	113	51
3	360	52
4	559	52
5	149	51
6	837	53
etc.		

Note that the probability of drawing, for example, 52 units sold on a given date is exactly equal to the probability shown in Table 17-6, since 350 numbers out of 1,000 were assigned to this event—daily sales of 52. Column 3 in Table 17-7 represents an artificially generated “history” of sales.

This history of sales could be used in a simulation model to study inventory control or the production or purchasing policy for the given product. It might also provide an input for a complex simulation model of the whole firm.

The procedure suggested above is appropriate for simulating distributions that are discrete. The appendix of this chapter discusses the simulation of continuous distributions.

Simulation of Complex Systems

Simulation was illustrated in the preceding pages by analysis of a simple queuing situation and by reference to Monte Carlo selection from probability distributions. But the great value of the simulation tool is in studying large complicated systems, which are too complex for mathematical analysis or simple judgment.

Consider as an example the operations of a barge line on the Ohio and Mississippi Rivers.⁷ The line operates tugs which pick up full barges of steel at the port of Pittsburgh and deliver the barges to downriver ports. At New Orleans, the tug turns around and picks up empty barges for the return trip. The barge line operates several tugs and hundreds of barges which are continually making the downriver trip and returning. There are many questions that management could ask about this system. These include: How should the tugs be scheduled (should they leave on a fixed schedule or wait until they have a full tow)? Should all tugs go through to New Orleans or should some turn around at an upstream port? How many tugs and barges should the firm own? Should the line seek general cargo for the return trip? All of these questions could be answered by building and analyzing a simulation model of the system.

The data needed for the simulation would include the probability distributions associated with (1) the availability of full barges of steel at Pittsburgh, (2) the distribution of the barges to the various destination ports and, (3) the distribution of turn-around time (for loading and unloading) at each port. Other factors, such as the time it takes a tug to go from one port to another and restrictions on the size of the tow, must also be included.

⁷ This example was suggested by the article “The Scheduling of a Barge Line” by G. O'Brien and R. Crane, *Operations Research*, Vol. 7, 1959, pp. 561-70.

By drawing random numbers and selecting values by Monte Carlo analysis for the probability distributions, we could simulate on paper (or in an electronic computer) the behavior of each tug as it obtains its load of barges and makes its downriver trip delivering the barges. So many barges are left at the first port, so many at the second, and so on. Similarly, the upriver trip could be handled by determining the number of empty barges available at each port. This system could be simulated under different schedules and with different amounts of equipment to determine the best policy for the barge line.

Simulation has also been applied to studies of inventory systems, to analyses of warehouse operations, to production scheduling, to studies of sales territories, to airline and railroad operations, to long-range financial planning, and in many other business areas. All these applications involve the building of a simulation model, usually on an electronic computer, to represent a real-world system. Different factors are then introduced into the model, and the results are analyzed. By this method the analyst can trace the effects of alternative policies and thus contribute to better decision-making.

SUMMARY

This chapter illustrates the use of certain probability models in business decision-making. Only a few representative models are included to demonstrate how probability analysis can be employed in specific situations.

The first model is concerned with a situation in which an individual must make a bid on some project in order to obtain a contract. The contractor estimates the probability distribution of the winning bid and then picks his bid so that it balances his profit, if he wins, with the probability that he will win. That is, the contractor selects a bid that maximizes his expected profit.

The inventory model involves the decision about how many units of a commodity to stock. The losses from overstocking and understocking are each a constant amount per unit. An optimal value of stock level can be determined from a formula involving the cumulative distribution of demand and the opportunity losses from overstocking and understocking. Goodwill costs associated with being out of stock and scrap allowances for resale value of unsold product can be included in the model if desired.

Waiting lines or *queues* develop at customer service stations at which the *arrivals* of customers or the times taken to *service* customers are variable amounts. If both arrivals and service completions follow a

Poisson distribution, the behavior of the queue can be described by a series of equations. These equations describe the probability that the queue contains a certain number of customers, as well as the expected length of the queue and the expected waiting time for a customer. The results may be used to design the service station to balance the costs of customer waiting with the cost of added facilities.

Simulation is a technique used to analyze complex business situations. A simulation model is built on paper or in a computer as an artificial representation of the real-world system. The simulation model is then operated as an approximation to the behavior of the business system through time.

In a simulation model, it is often necessary to represent the behavior of a random variable. This may be done by *Monte Carlo* analysis, the artificial construction of a history of random occurrences, based upon a probability distribution.

APPENDIX: THE MONTE CARLO METHOD FOR CONTINUOUS DISTRIBUTIONS

When we are trying to obtain random drawings from a continuous distribution, the analysis is basically the same as for discrete distributions. The first step is to determine the cumulative probability distribution for the random variable involved. As an example, let us return to the queuing illustration on pages 408 to 410 in which the history of arrivals was given. Suppose, instead, that no past data were available but arrivals were expected to occur at random with a Poisson distribution and an arrival rate of 18 per hour or 0.3 per minute. When the arrivals are Poisson-distributed, the random variable *time between arrivals* has a continuous distribution known as the *exponential distribution*.⁸ Since the arrival rate is 0.3 per minute, the mean time between arrivals is $1/0.3$ or $3\frac{1}{3}$ minutes. Then t , the time between arrivals, can be described by the cumulative distribution shown in Chart 17-4. The chart shows the probability that the time between arrivals will be equal to or less than the indicated number of minutes. For example, the probability is approximately 0.60 that an arrival will occur within 3 minutes of the previous arrival.

⁸ The exponential distribution has the following form:

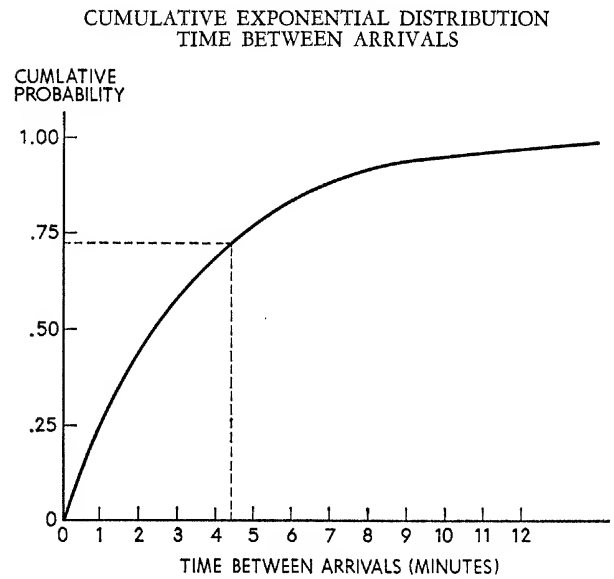
$$g(t) = \alpha e^{-\alpha t}$$

where t is the random variable time between arrivals; $g(t)$ is the probability function of t ; $\alpha = 0.3 =$ arrival rate per minute (the reciprocal of average time between arrivals); and e is the constant 2.718 The cumulative distribution $G(t)$ of t has the following form:

$$G(t) = 1 - e^{-0.3t}$$

This is the curve plotted in Chart 17-4.

Chart 17-4



Note that for every value of the cumulative probability there is a corresponding value of t . Also, the cumulative probability ranges from 0 to 1.0. By selecting a random number between 0 and 1.0, we can find an associated value of t . Thus, if we selected the random number .73 or 0.73, the associated value of t is 4.3, as shown by the dashed lines in Chart 17-4. By repeatedly drawing random numbers between 0 and 1.0, we can generate a whole series of values for t , the time between arrivals.

Table 17-8

SIMULATING A HISTORY OF ARRIVALS
BY RANDOM NUMBERS AND A PROBABILITY DISTRIBUTION

Arrival Number	Random Number	Random Time between Arrivals from Chart 17-4	Time of Arrival = Time of Previous Arrival + Time between Arrivals
0			0:00.0
1	0.73	4.3	0:04.3
2	0.04	0.1	0:04.4
3	0.97	11.3	0:15.7
4	0.38	1.6	0:17.3
5	0.68	3.8	0:21.1
6	0.26	1.0	0:22.1
etc.			

This is done for a few values in Table 17-8, which turn out, by intention, to be the same arrivals as in the first part of Table 17-5.

PROBLEMS

1. A contractor is about to bid on a certain job that he estimates will cost him \$80,000. His judgment about the winning bid is expressed as a normal distribution with mean \$100,000 and standard deviation \$10,000. What bid should he make to maximize his expected profit?
2. A large corporation plans to purchase a fleet of 1,000 automobiles for its salesmen and has asked for bids. You collect data on the several similar bidding situations in the recent past. The winning bid has varied depending upon the style of car and the type of equipment desired. In each case, however, you determine the difference between your cost estimate per automobile and the winning low bid. These are shown in the table.

Difference between Winning Low Bid and Your Company Cost Estimate	Frequency
0 to \$ 50.00	2
\$ 50.01 to 100.00	6
100.01 to 150.00	4
150.01 to 200.00	3
200.01 to 250.00	1
250.01 to 300.00	2
300.01 to 400.00	0
400.01 to 500.00	1
500.01 to 600.00	1
Total	20

How much above your cost estimate should you bid in order to maximize your expected profit?

3. The city of Zenith has called for bids on a new generator for its municipal electric company. The generator is to be constructed to specifications determined by Zenith's power engineers. You expect that approximately three firms will submit bids, and the generator will be purchased from the lowest bidder.

As president of a small firm, Ridgway Dynamo and Engine Company, you wish to be as careful as possible in making your bid. Your engineers estimate that the variable cost of manufacturing the generator (cost of labor, materials, equipment) is \$180,000. In addition, 30 percent (or \$54,000) is to be added to this cost for overhead and fixed costs, making a total estimated cost of \$234,000. You have sufficient excess capacity to manufacture the generator and, in fact, will be forced to lay off part of your work force if you do not get the bid.

As a means of determining your bid, you collect the information shown in the table on the last 24 bids on electric generators.

Job	Cost Estimate (\$000)			Ridgway Bid, (\$000)	Winning Bid (\$000)	Winning Bidder
	Variable Cost	Overhead	Total Cost			
1	84.0	16.8	100.8	123.7	123.7	Ridgway
2	232.4	46.5	278.9	362.1	357.5	Westinghouse
3	187.5	37.5	225.0	275.0	270.6	Westinghouse
4	68.2	13.6	81.8	115.6	111.1	G.E.
5	147.2	29.4	176.6	212.6	208.6	Elliott
6	240.0	48.0	288.0	362.4	360.0	G.E.
7	328.2	98.5	426.7	560.2	554.0	Westinghouse
8	415.7	124.7	540.4	594.1	564.1	Westinghouse
9	98.3	29.5	127.8	146.2	138.3	G.E.
10	62.7	25.1	87.8	100.2	95.7	Westinghouse
11	171.6	68.6	240.2	284.2	222.0	G.E.
12	198.0	79.2	277.2	310.1	262.1	Westinghouse
13	203.1	71.1	274.2	282.8	282.8	Ridgway
14	110.0	38.5	148.5	178.2	149.8	Westinghouse
15	167.2	58.5	225.7	276.8	259.7	Elliott
16	214.0	53.5	267.5	340.0	320.0	Elliott
17	308.9	77.2	386.1	465.0	451.5	G.E.
18	224.5	56.1	280.6	345.4	345.4	Ridgway
19	180.0	36.0	216.0	281.5	251.5	Westinghouse
20	241.2	48.2	289.4	342.1	336.8	G.E.
21	164.8	33.0	197.8	245.2	233.6	G.E.
22	142.4	42.7	185.1	218.6	192.5	Westinghouse
23	200.0	60.0	260.0	285.0	250.0	Westinghouse
24	178.2	53.5	231.7	310.2	289.4	G.E.

What bid do you think should be made in this situation? Why?

4. Refer to Problems 5 and 6 at the end of Chapter 9. Find the solutions to these problems using the inventory model discussed in Chapter 17.
5. Refer to Problem 8 at the end of Chapter 9. Why can this problem not be solved using the inventory model of Chapter 17?
6. A vendor buys a certain product for \$1.18 and sells it for \$1.98. Any items unsold at the end of the period are disposed of at a price of 64 cents. The demand for the product follows this distribution:

Demand X	Probability $P(X)$
180	0.15
190	0.40
200	0.20
210	0.15
220	0.05
230	0.03
240	0.02
Total.....1.00	

- a) Assuming no goodwill loss associated with being out of stock, how many should be purchased?
- b) If the goodwill loss is 50 cents each, how many should be purchased?

7. Refer to Problem 14 at the end of Chapter 15.
- Suppose you had expressed your uncertainty about the number attending by a normal distribution with mean equal to 300 and standard deviation of 80. Using the same *cost* information and excluding any sample information, how many dinners should you order?
 - Assuming the normal prior distribution of part *a*, suppose that a sample of 100 persons are contacted with the result that 20 indicate they will attend. Using this information, how many dinners should you order? (*Hint*: Use the normal approximation to the binomial and ignore the finite population correction factor. Revise the normal prior probabilities as described in Chapter 16.)
8. Demand for a certain product is known to be approximately normal with mean 100 units and standard deviation 20 units. The product costs \$1 each and sells for \$2.40. Items unsold at the end of the period have no value.
- If there is no goodwill loss associated with being out of stock when a customer wants a unit, what is the optimal stock q_* ?
 - The manager in charge of the inventory for this product has traditionally stocked 120 units. When shown the answer to *a* above, he states that he has incorporated a goodwill loss for being out of stock. What is the implicit goodwill loss associated with the inventory policy of 120 units?
9. A buyer in the toy department for a group of department stores must place his order for a certain toy for the Christmas season by late spring. The toy is a plastic model truck which has a retail price of \$14.98. In quantity lots, the toy will cost the store \$7.28 to purchase.

The buyer was undecided about the quantity to order not only because of uncertainty about whether the Christmas season would be "good" or "poor" but also because of uncertainty about the appeal of the particular toy. He knew that certain toys became favorites and the stores could sell virtually all they could buy, while other toys were less popular and sold only a few. The buyer, after some thought, expressed his judgment in the form of the following bimodal probability distribution:

Sales, Units	Probability	Sales, Units	Probability
100	0.03	220	0.02
120	0.15	240	0.05
140	0.10	260	0.15
160	0.05	280	0.25
180	0.03	300	0.10
200	0.02	320	0.05
Total.....		1.00	

Assume that units unsold over the Christmas season must be sold to dealers for \$5.14. In addition, there is a handling cost of \$1 for each leftover unit.

- If no goodwill loss is associated with being out of the particular toy, how many should the buyer order? What is the expected profit?

- b) Suppose that when a customer comes in to buy this particular toy and it is not available he will go elsewhere and buy all his toys. Suppose the lost profit from thus losing a customer is \$10 per customer. Under this assumption, how many should be ordered?
- c) Suppose that, instead of the situation described in *b* above, a customer who cannot buy the particular toy spends his money on other toys with a profit of \$6 per customer. Under this assumption, how many should he order?
- d) Suppose that 20 percent of the customers are as described in *b* and 80 percent are as described in *c*. How many toys should be ordered?
10. The Fox Photo Company is a mail-order firm specializing in 24-hour service on developing negatives and making prints. The general policy is that orders arriving in the morning mail must be finished and in the outgoing mail before the midnight mail pickup. This has usually involved little difficulty. Six full-time technicians work an eight-hour day from 8 A.M. to 5 P.M. and are paid at a rate of \$4 per hour (including fringe benefits). These technicians can process an average of 5 orders an hour. When, on occasion, more than about 240 orders arrived in a given day, one or more of the men work overtime at a rate of \$6 per hour.
- Fox Photo has recently bought out a competitor in the same community and plans to consolidate operations. Mr. Fox is undecided, however, on how many technicians to add to the six he now employs. By adding together the past order data of his competitor to his own, Mr. Fox has the following frequency data to ponder:

Number of Incoming Orders	Fraction of Days
Under 220	0.03
220-239	0.03
240-259	0.09
260-279	0.16
280-299	0.18
300-319	0.20
320-339	0.15
340-359	0.10
360-379	0.05
380 and above	0.01
Total.....	1.00

One of the technicians at Fox Photo was taking a night course in statistics at a local college and tried his hand at analyzing the above data. After a couple of evenings work he told Mr. Fox that the data closely fit a normal distribution with mean 300 and standard deviation 40. But the technician was unable to answer the question of how many technicians to employ.

- a) How many additional technicians should Mr. Fox employ? What is his expected cost?
- b) What additional factors should be included in making this decision?

11. A new branch of the National Bank is under construction. One drive-in window is planned. The branch manager is worried because current plans allow room only for a line of three cars at the window (the car being serviced plus two cars waiting). The manager feels that he may lose customers who would otherwise enter the line but cannot because of space limitations.

Suppose that customers are expected to arrive at a rate of 10 per hour and the average service time is 3 minutes.

- a) If there were unlimited space, what would be the probability of more than two cars in the queue (excluding the car at the window)? What would be the average waiting time?
 - b) If banking hours are from 10 to 3 (5 hours), what is an upper limit on the number of customers per day that would be turned away because of the space limitations?
12. The Lakes Ore Company (LOC) wished to expand the number of shipments of iron ore across the lakes. However, the dock facilities at the port were inadequate and new equipment would be needed. During the 1968 season, LOC expected to ship approximately 90 shiploads of ore during the 180 days of peak operations—April 15 to October 12.

LOC had dock space for only one ship and wished to minimize waiting time since a ship's operating cost was \$200 per day.

Two different methods of unloading ships were under consideration. One method, *A*, used considerable manual labor, and required an average of $1\frac{1}{4}$ days to unload a ship. This method would cost \$500 per ship unloaded. Method *B*, on the other hand, was considerably more mechanized and cost \$800 per ship unloaded. However, ships could be unloaded at a rate of one a day on the average.

Assume that weather and other factors cause the ships to arrive in port in a random fashion.

- a) Suppose that service completions were also random (i.e., Poisson distributed); which method (*A* or *B*) should be used for unloading the ships?
 - b) Is the Poisson assumption likely to be reasonable in this case?
 - c) Suppose that the time taken to unload a ship using Method *A* was always *exactly* $1\frac{1}{4}$ days and the time for Method *B* exactly 1 day, how would this modify your answer to part *a* above?
13. Mr. Jones is the reservation clerk at the New York office of Cross America Airlines. Jones has a long-standing argument with his supervisor about the installation of a new reservation system that would speed his work. Jones has argued that many callers for reservations must wait until he is free and that there is a "goodwill" cost of such a wait. Jones contends that waiting detracts from the image of the airline as efficient, friendly, and personal. Further, Jones feels that some who are forced to wait will fly on competitive airlines.

Mr. Smith, the vice-president of Reservation Services, does not agree. He points to the fact that Jones is idle a good part of the time and that the

new equipment would be costly—it would cost approximately \$90 to operate for an eight-hour day.

The argument has gone unresolved. However, one day Mr. Jones attended a management development dinner at which a professor from a leading School of Business spoke on the applications of waiting-line theory to business problems. Convinced that he could marshal this “scientific” technique behind his argument with Smith, Jones eagerly began to collect data, as shown in the table.

CROSS AMERICA AIRLINES NUMBER OF CALLS ARRIVING PER 5-MINUTE PERIOD 1,000 PERIODS—NEW YORK OFFICE	
Number of Calls in a Given 5-Minute Period	Number of Periods Event Occurred
0	350
1	370
2	190
3	60
4	20
5 and up	10
Total	1,000
Average number of calls arriving per 5-minute period = 1	

Jones also kept data on how long it took him to service a caller. For the 1,000 calls in the table, it took Jones, on the average, $2\frac{1}{2}$ minutes for each customer. Many calls, of course, took much less time as, for example, when the caller merely wanted the arrival time of a certain flight. Occasionally, it took as much as 10 minutes to help a customer if it involved complicated schedule arrangements.

Mr. Jones further proceeded to make some cost estimate for the “good-will” lost from a customer waiting. He felt that 40 cents per minute was a reasonable figure, but when he suggested this amount to Smith, it was received with some disdain. Smith felt that 10 cents per minute was the “outside (maximum) limit” for such a cost, provided that persons did not have to wait longer than 4 or 5 minutes.

Jones was uncertain how to proceed further in analyzing his problem. He felt that the proper “scientific” (waiting line) solution would show that the new equipment would save money.

- a) What can you tell Jones about the value of the new equipment?
- b) What other alternatives might be considered by Cross America?
- c) Would you expect the average number of calls to be the same over the period of one day? How would this affect the analysis?

14. The tool crib in a certain factory is a room where special tools, jigs, and other equipment are stored for general use by mechanics. An attendant signs the equipment in and out as the mechanics request it or return it. The production foreman has been concerned because occasionally many mechanics line up at the tool crib with considerable waiting and lost production time.

The clerk in charge of the tool crib suggests that an assistant be hired to help. The assistant would help find equipment and thus speed up the service to the mechanics. (It would still be a one-channel operation, however, with the clerk checking the equipment in and out.) The assistant would be paid \$1.85 per hour plus fringe benefits of 35 cents per hour. He would work one shift (8 hours per day).

A check was made to determine how many mechanics came to the tool crib. From the records it was determined that an average of 15 mechanics came to the tool crib per hour. A study was undertaken to determine how long it took the clerk to wait on a mechanic with a resulting estimate of 2.4 minutes per mechanic on the average. It was estimated that the clerk could wait on 30 mechanics per hour if he had a helper.

Mechanics are paid at a rate of \$5 per hour plus 40 cents in fringe benefits. Assume that arrivals and service times at the tool crib are random.

- a) What is the probability that a mechanic will have at least some wait under the present system? If a helper is hired?
 - b) What is the average wait for a mechanic under the present system? With the helper?
 - c) Should the helper be hired?
15. Refer to Problem 12 above (Lakes Ore Company). Simulate for 200 days the situation described in part (c). That is, arrivals of ships follow a Poisson distribution and service (unloading) times are exactly $1\frac{1}{4}$ and 1 days for the two alternatives. For simplicity, assume that ships arriving on a given day all come in at a certain time, for example, 8AM. Compare your answer with that obtained in Problem 12 (a).
 16. Refer to Problems 12 and 15 above. Suppose Lakes Ore Company could build a second dock so that two ships could be handled simultaneously. The cost of the second dock would be \$100 per day. The unloading method at each dock would be the more manual type, involving a cost of \$500 per ship unloaded, and an unloading time of $1\frac{1}{4}$ days. Simulate operations over 200 days and compare the cost of this alternative to that of Problem 15.
 17. An investor with \$300 is considering the purchase of three stocks, *A*, *B*, and *C*, each selling for \$100 a share. He attaches the probabilities shown in the table below to the value (dividends plus market price) of the stocks at the end of one year.

Value at End of Year	Stock <i>A</i> Probability	Stock <i>B</i> Probability	Stock <i>C</i> Probability
\$ 90		0.20	0.30
100	0.50	0.20	0.10
110	0.40	0.20	0.10
120	0.10	0.20	0.10
130		0.20	0.40
Totals	1.00	1.00	1.00

- a) Suppose the investor wishes to buy one share of each stock. Assume that the stocks are independent (i.e., the year-end value of one is not related to the value of any other). Use Monte Carlo analysis to estimate the probability distribution associated with the value of the portfolio of three stocks at year's end. Calculate the mean and variance of this distribution.
- b) Compare the mean and variance of the portfolio obtained in *a* above with the mean and variance of the alternatives of buying three shares of Stock *A*, or three shares of Stock *B*, or three shares of Stock *C*.
18. Refer to Problem 17. Suppose that a fourth stock, Stock *D*, is available at a price of \$100 per share and is unrelated to Stocks *A* and *B* but is related to Stock *C* as shown by the probabilities in the table.

Value of Stock <i>C</i> at End of Year	Value of Stock <i>D</i> at End of Year					Totals
	\$90	\$100	\$110	\$120	\$130	
\$ 90				0.20	0.10	0.30
100			0.10			0.10
110			0.10			0.10
120			0.10			0.10
130	0.20	0.10	0.10			0.40
Totals.....	\$0.20	\$0.10	\$0.40	\$0.20	\$0.10	\$1.00

- a) By the use of Monte Carlo analysis, estimate the distribution of year-end value of a portfolio composed of Stocks *A*, *C* and *D*. Determine the expected value and variance of this distribution.
- b) By the use of Monte Carlo analysis, estimate the distribution of year-end value of a portfolio composed of Stocks *B*, *C*, and *D*. Determine the expected value and variance of this distribution.
- c) A portfolio of stocks is defined as "efficient" if there is no other portfolio with the same variance having a higher expected value—or, alternatively, if there is no other portfolio with the same expected value having lower variance. Which of the portfolios considered in Problems 17 and 18 are efficient in this sense? Which are inefficient? (Note: Only the portfolios *AAA*, *BBB*, *CCC*, *ABC*, *ACD*, and *BCD* have been considered. There are, of course, others such as *AAB*—two shares of Stock *A* and one of *B* etc. For simplicity, ignore these possibilities.)
19. In the typical "two-bin" inventory situation, an order for replenishment is made when the stock level reaches an amount *b*. The order is made for an amount *q*, called the order quantity. It takes a certain number of days, called the "lead time," until the order comes in. During this lead time if sales exceed the order level *b*, a stock-out condition occurs and sales are lost with cost *k*. It generally costs a certain amount *c_o* to place an order and a certain amount *c_h* to hold one unit of inventory in stock over a period of time (say, a year).

In the usual situation the probability distribution of demand for the

product is given as well as the lead time. The constants c_o , c_h , and k are estimated. Then the values for order level b and the order quantity q must be determined to minimize cost over a period of time.

One method of dealing with this problem is to simulate the inventory system for different values of b and q and to use the results of the simulations to determine good values for b and q .

Suppose that the daily demand for a certain product is as shown in the table.

Daily Demand, Units	Probability
0	0.10
1	0.30
2	0.20
3	0.10
4	0.10
5	0.10
6	0.05
7	0.05
Total.....	1.00

The lead time (time from when an order is placed until it comes in) is 20 days. Suppose that cost of being out of stock is $k = \$3$ per unit for each stock-out. The cost of placing an order is $c_o = \$10$, and the cost of holding one unit of inventory is 50 cents per month (30 days).

- Assume that the order quantity q is fixed at 55 units. Simulate 300 days operations for each of 3 different values of b , the stock level. Estimate the cost for each system. Which value of b is best? Do you think the "best" value of b is greater than or less than the value you obtained?
- Select three different sets of values for q and b . Simulate 300 days operations for each set and estimate the cost of the inventory system for each set. Which set gave the lowest cost?

SELECTED READINGS

EDELMAN, FRANZ. "Art and Science of Competitive Bidding," *Harvard Business Review*, (July–August, 1965).

An application of the bidding model discussed in this chapter.

GOETZ, BILLY E. *Quantitative Methods—A Survey and Guide for Managers*. New York: McGraw-Hill, 1965.

This text covers a wide range of models useful in decision-making, including many probability models. Chapter 11 treats simulation in detail.

KEMENY, J. G.; SCHLEIFER, A., JR.; SNELL, J. L.; and THOMPSON, G. L. *Finite Mathematics with Business Applications*. Englewood Cliffs, New Jersey: Prentice-Hall, 1962.

The authors treat Monte Carlo simulation with applications to queuing and decision-making situations on pp. 199–228.

LEE, ALEC M. *Applied Queueing Theory*. London: Macmillan, 1966.

An eminently useful book on the practical aspects of solving queuing

problems. Earlier chapters survey queuing models. Later chapters discuss specific examples.

McMILLAN, C., and GONZALEZ, R. F. *Systems Analysis*, Homewood, Illinois: Richard D. Irwin, 1965.

This book is concerned with the development of computer models for business systems. Probability models are emphasized. In particular, Chapter 6 treats Monte Carlo simulation; Chapters 7 and 9 deal with queuing models. In general, it is a good source for a more detailed treatment of the topics in this chapter, at a level comparable to this text.

MORGENTHAUER, GEORGE W. "The Theory and Application of Simulation in Operations Research" in Russell L. Ackoff, ed., *Progress in Operations Research*, Vol. 1. New York: John Wiley, 1961, pp. 363-420.

A good discussion of simulation and Monte Carlo analysis, including a section dealing with specific applications of simulation to business problems. The bibliography gives a detailed list of published material on simulation.

SASIENI, M.; YASPAN, A.; and FRIEDMAN, L. *Operations Research Methods and Problems*. New York: John Wiley, 1959.

Chapter 4 is concerned with inventory models. Chapter 6 deals with queuing models and simulation. A derivation of Equations 3 through 8 is included.

SCHLAIFER, R. *Probability and Statistics for Business Decisions*. New York: McGraw-Hill, 1959.

Chapter 4 deals with simple inventory problems; Chapters 19 and 20, with queuing and Monte Carlo analysis.

18. INDEX NUMBERS

INDEX NUMBERS express the *relative* changes in a variable compared with some base, which is taken as 100.¹ The variable may be a single series, such as electric power production, or an aggregate, such as a group of common stock prices. The index number usually represents a sample of such a group. The changes measured may be those occurring over a period of time or those between one place and another.

Many aspects of modern business are described by the use of index numbers. Both government and private agencies are devoting increasing efforts to the construction of index numbers as aids in management and in the interpretation of changes in general economic life. Many businesses use a variety of index numbers for their own internal administrative purposes. Certain statistical publications, notably the *Survey of Current Business*,² *Economic Indicators*, *Business Cycle Developments*, *Federal Reserve Bulletin*, and the *Statistics* bulletin of Standard and Poor's Corporation, contain hundreds of economic time series expressed in index number form.

Statistical ingenuity has developed an almost encyclopedic list of uses of business indicators. The most important of these are (1) measures of the economic well-being of the economy, a geographic area, an industry, or a specific business; (2) comparisons of related series for administrative purposes; (3) the use of price indexes as deflators to express a value series in constant dollars; (4) the use of price indexes as escalators in wage and other contracts; (5) specific guides or "triggers" for the

¹ The term "index" is sometimes applied to a business indicator expressed in any unit. Thus, pig-iron production in tons may be referred to as an "index" of business activity. In this chapter, however, the term "index number" or "index" refers specifically to a ratio having some base as 100, or to a series of such ratios.

² Summary descriptions of 2,500 series may be found in the footnote references of the biennial *Business Statistics* supplement to the *Survey of Current Business*.

initiation of administrative business or government actions; and (6) the basis or orientation for forecasting.

ADVANTAGES OF INDEX NUMBERS

Index numbers are widely used because they have the following important advantages, in contrast with actual data:

1. They provide a simple method of comparing changes from time to time or from place to place. It is easy to compare 83 cents for a pound of ham with 25 cents for a quart of milk, but it is not so easy to compare price changes in the two articles over a period of time. Index numbers of the ham and milk prices would indicate the relative change in each price from some given price and which of the two prices had shown the greater change (see Table 18-4). As the number of items increases, this advantage becomes even more apparent.

2. Index numbers facilitate comparison of changes in series of data expressed in a variety of units—for example, dollars, tons, or gallons. Data pertaining to production, sales, inventories, costs, or other aspects of business may also be put into index number form and then compared.

3. They make possible the construction of composites that represent in a single figure some overall measure of business. This simplifies comparisons with other types of data. In January 1967, the U.S. Bureau of Labor Statistics Index of Wholesale Prices stood at 106.2. This single figure indicates the average relation of prices in January 1967 to prices in 1957-1959, the base period for this index, taken as 100. That is, it took \$10.62 to buy the same amount of specified goods as could have been bought for \$10 in 1957-59.

Even series expressed in different types of units sometimes can be combined into a meaningful aggregate, provided the combinations make sense. Many examples of such combinations appear throughout this chapter.

4. They describe the typical seasonal patterns of business. The annual peak in department store sales, for instance, regularly occurs in December, while sales of soft drinks are greater in midsummer. These "indexes of seasonal variation" are described in Chapter 20.

KINDS OF INDEX NUMBERS

An examination of any journal of business statistics will reveal many different index numbers which describe changes in various aspects of business and economics. These index numbers may be classified as (1) price indexes, (2) quantity indexes, and (3) value indexes. Some of the most commonly used indexes of these three types, and their principal

Table 18-1

SOURCES OF COMMONLY USED INDEXES*

Name of Index	Prepared by	Frequency of Publication	Published Regularly in
A. PRICE INDEXES			
1. Consumer Price Index	U.S. Bureau of Labor Statistics	M	SCB, FRB, MLR, <i>Business Week</i> , S&P, <i>Ec. Ind.</i> , NICB
2. Wholesale Price Index	U.S. Bureau of Labor Statistics	W, M	SCB, FRB, MLR, NICB
3. Spot Market Prices of 22 Basic Commodities	U.S. Bureau of Labor Statistics	D, M	Barron's, C&FC, S&P, <i>Ec. Ind.</i>
4. Construction Cost Indexes	American Appraisal Co.	M	Barron's, SCB, S&P
5. Stock Price Averages	Dow-Jones & Co.	H, D, W, M	SCB, Barron's, S&P, C&FC
6. Stock Price Index, 500 Stocks	Standard and Poor's Corp.	H, D, W, M	SCB, FRB, S&P, <i>Ec. Ind.</i> , <i>Business Week</i>
B. QUANTITY INDEXES			
1. Industrial Production	Federal Reserve Board	M	SCB, FRB, S&P, <i>Ec. Ind.</i> , NICB
2. Production and Trade	Barron's	W	Barron's
3. Steel Production	American Iron and Steel Institute	W, M	SCB, Barron's, C&FC
4. Business Failures	Dun and Bradstreet	W, M	Barron's, C&FC
C. VALUE INDEXES			
1. Gross National Product	U.S. Department of Commerce	Q	SCB, FRB, S&P, <i>Ec. Ind.</i> , NICB
2. Manufacturing Production-Worker Payrolls	U.S. Bureau of Labor Statistics	M	SCB, FRB, MLR, S&P, C&FC
3. Construction Contracts Awarded (Value)	F. W. Dodge Corp.	M	SCB, FRB, <i>Ec. Ind.</i>
4. Measure of Personal Income (by states)	<i>Business Week</i>	M	<i>Business Week</i>

* Abbreviations:

H—hourly or shorter intervals; D—daily; W—weekly; M—monthly; Q—quarterly.

SCB—*Survey of Current Business* (and weekly supplement)FRB—*Federal Reserve Bulletin*MLR—*Monthly Labor Review*C&FC—*Commercial and Financial Chronicle*S&P—Standard and Poor's *Trade and Securities Statistics**Ec. Ind.*—President's Council of Economic Advisers, *Economic Indicators*NICB—National Industrial Conference Board, *Selected Business Indicators*

sources, are listed in Table 18-1. Most of these, but not all, are expressed in relative form.

Price Indexes

Some of the best-known indexes are those dealing with prices. Prices have been of widespread interest for centuries as sensitive barometers of industry and trade.

The necessary data for price index numbers arise from the exchange of commodities (1) at different stages of production—raw materials, semifinished goods, and completely fabricated products; (2) at several levels of distribution—industrial, wholesale, and retail; and (3) for a variety of groups of items—consumers' goods, producers' goods, stocks and bonds, durable and nondurable goods.

A *purchasing power index* is the reciprocal of a price index, when both indexes are expressed as ratios with base 1 rather than 100. Taking the wholesale price index of 106.2 for January 1967 as 1.062, its reciprocal is $1/1.062 = 0.942$, so the corresponding purchasing power index (with base 100) is 94.2. This means that for every dollar's worth of goods one could buy at 1957–1959 wholesale prices, one could buy 94.2 cents' worth in January 1967. Hence, the January 1967 dollar was worth only 94.2 cents in comparison with the 1957–1959 dollar.

Quantity Indexes

Quantity indexes measure the physical volume of production, construction, or employment. They are computed for (1) industry in general, (2) specific industries, or (3) specific operations or stages of production or distribution. The data may represent the country as a whole or local trading areas.

Because of the nature of the data, quantity index numbers are frequently less reliable than those based on dollar figures. Historically, business records were designed to include chiefly those aspects of business which could be expressed in monetary units and, consequently, data in physical units for extended periods of time are difficult to obtain.

Value Indexes

Value indexes show the total dollar volume of income, payrolls, sales, and the like. Value is the result of multiplying quantity by price; index numbers of value therefore reflect changes in both quantity and price. The gross national product estimates of the U.S. Department of Commerce are constructed much like other value indexes, but they are expressed in billions of dollars rather than as percents of a base to avoid the "aura of normality" attached to a base period.

It will be noted that the *New York Times* and *Barron's* indexes of general business activity measure physical volume changes, such as tons of steel and kilowatts of electricity produced, while many regional indexes measure dollar volume, such as factory payrolls and department store sales. Some regional business barometers even combine quantity and value measures, but these indexes are more difficult to interpret.

BASIC METHODS OF CONSTRUCTING INDEX NUMBERS

Simple Index Numbers

A simple index number is constructed from a single series of data which either extends over a period of time or simultaneously represents several different locations. In constructing such an index number, one particular period or place is selected as the base and the item for this base is taken as 100. The other items in the series are then expressed as percents of this base. A simple index is frequently called a price *relative*, quantity relative, or value relative.

As an example of a quantity relative, an airline executive may wish to compare the changes in air and automobile travel from 1960 to 1965. Since the volume of intercity automobile passenger-miles traveled is over 15 times that of air travel, the executive's purpose would not be accomplished by comparing the changes in actual passenger-miles. The two series can be more easily compared if they are expressed as percentages of passenger-miles traveled in the same base period—say, 1960.

The construction of these simple indexes or quantity relatives is shown in Table 18-2. The three steps are (1) choose the base period (1960); (2) divide the travel figure each year by the base figure;³ and (3) multiply the result by 100 (i.e., move the decimal point two places to the right) to express it as a percent or index number. An index number is written just as a percent, except that the percent sign (%) is not used. Thus, the 1965 index for air travel is $51.9 \div 30.6 \times 100 = 170$.

This index means that air travel in 1965 was 170 percent of its 1960 volume, an increase of 70 percent. Hence, while automobile travel had increased more than air travel in passenger-miles during this period (136 billion versus 21.3 billion), its *relative* increase was only 20 percent, compared with 70 percent for air travel.

The increase in the air travel index from 1964 to 1965 was 26 index points, but this is not 26 percent because the base is 144, not 100. The *percentage* increase was $26 \div 144 = 18$ percent.

A simple index can be computed for any single series of data, such as the price of General Motors stock or a department store's sales. Statisti-

³ Whenever it is necessary to divide a series by a constant divisor, as in this instance, it is usually easier to use the reciprocal of the divisor as a fixed multiplier. In this example, air passenger-miles can be simply multiplied by the reciprocal of 30.6 (found in Appendix C) $\times 100 = 3.27$. This figure can be kept in the calculating machine without change throughout the entire computation, thus saving time and reducing the likelihood of error.

Table 18-2

SIMPLE INDEX NUMBERS OF AIR TRAVEL
AND INTERCITY AUTOMOBILE TRAVEL
IN THE UNITED STATES, 1960-1965

Year	Passenger-Miles (Billions)		Index (1960 = 100)	
	Air Travel	Auto Travel	Air Travel	Auto Travel
1960	30.6	681	100	100
1961	31.1	692	102	102
1962	33.6	720	110	106
1963	38.5	748	126	110
1964	44.1	783	144	115
1965	51.9	817*	170	120

* Estimated.

SOURCE: *Air Transport Facts and Figures*, 1966, p. 33.

cal source books include many indexes of this type. The Bureau of Labor Statistics, for example, publishes monthly price relatives for each of about 2,200 commodities, as an aid in comparing individual price changes, in addition to its composite wholesale price indexes.⁴

Composite Index Numbers

Most index numbers in common use are composites. They are constructed according to the principles just described for simple indexes, but they combine several different sets of data. In the following pages, two basic methods of constructing composite index numbers are described: (1) the average of relatives index and (2) the aggregative index. Formulas for both types of indexes are presented on page 437, but it is not necessary to memorize them to understand the procedure involved.

Necessity of Weights. Whenever prices or other data are combined in an index number, the relative importance of each must be taken into account by assigning proper weights to each item. This is necessary because, in reality, no composite index is unweighted. If a set of weights is not explicitly applied, each element of the index *automatically* (or implicitly) receives *some* weight. For example, if unit prices of various foods are being added together in the preparation of a composite consumer price index, a given relative change in a higher-priced item, such as a pound of ham, will influence the total more than will the same relative change in a lower-priced item, such as a quart of milk. Milk, however, should really be weighted more heavily because

⁴ See U.S. Bureau of Labor Statistics, *Wholesale Prices and Price Indexes*, 1962, Bulletin No. 1411 (July 1965).

people consume more; so a system of weights must be used in order to give milk its proper importance in the index. A composite index is thus a *weighted average*⁵ of its components.

Average of Relatives Method. Many methods of constructing index numbers have been tried, but the average of relatives method is now used in most leading indexes, such as the Federal Reserve Board's index of industrial production and the Bureau of Labor Statistics' wholesale price indexes. In this method the individual series of price or quantity data are expressed as simple indexes, which are then multiplied by fixed *dollar value weights* and totaled to yield the composite index.

To illustrate the construction of a *quantity* index, consider a manufacturer of light-weight airplane luggage and specially fitted car-top luggage for automobiles. About two thirds of his sales are typically airplane luggage and one-third is car-top luggage. He wishes to construct a composite index of air and automobile travel and project it into the future as a measure of the potential market for his products. The method is illustrated in Table 18-3. The steps are as follows:

1. Express each individual series as a simple index or relative, by dividing through by the base value. This step is described above. (Columns 1-3 in Table 18-3 are taken from Table 18-2.)
2. Select a dollar-value weight for each series as a measure of its importance in the base year or some other typical period. Divide these weights by their total to express them as *relative* weights whose sum equals 1. In this case the relative importance of air and auto travel *to the manufacturer* is measured by the proportion of his dollar sales that go to each industry— $\frac{2}{3}$ and $\frac{1}{3}$, respectively. As a more general example, the Federal Reserve Board weights its component indexes of manufacturing output by "value added by manufacture," from the Census of Manufactures, expressed as percents of the total weight.
3. Multiply the simple indexes by the relative weights to obtain the weighted indexes (Table 18-3, columns 4 and 5).
4. Add the weighted indexes to obtain the composite index (column 6). This must equal 100 in the base year, since the simple indexes

⁵ The weighted arithmetic mean is used almost universally in computing index numbers, although the weighted geometric mean is theoretically superior for averaging relatives, particularly since they tend to follow a logarithmic normal distribution, with a zero lower limit and infinite upper limit. The geometric mean also minimizes the influence of extremely large relatives, which may distort the arithmetic mean of a small number of items. Nevertheless, the arithmetic mean is used because it is easier to compute and easier to understand than the geometric mean. Also, an arithmetic price index represents changes in the total cost of a bill of goods more accurately than a geometric index, which reflects the average ratios of change in price. That is, the arithmetic mean makes more sense in this connection.

equal 100 and the weights total 1. (If the value weights are *not* adjusted to total 1, the sum of the weighted indexes can be divided through by its base-year value to obtain the same values as in column 6 of the table.)

Table 18-3

CONSTRUCTION OF COMPOSITE INDEX
OF AIR AND AUTOMOBILE TRAVEL
BY AVERAGE OF RELATIVES METHOD
(1960 = 100)

Year (1)	SIMPLE INDEX (1960 = 100)		WEIGHTED INDEX		COMPOSITE INDEX
	Air Travel (2)	Auto Travel (3)	Air Travel (Column $2 \times \frac{2}{3}$) (4)	Auto Travel (Column $3 \times \frac{1}{3}$) (5)	Air and Auto Travel (Columns $4 + 5$) (6)
1960	100	100	67	33	100
1961	102	102	68	34	102
1962	110	106	73	35	108
1963	126	110	84	37	121
1964	144	115	96	38	134
1965	170	120	113	40	153

SOURCE: Table 18-2.

The composite index provides the manufacturer with a summary measure of potential demand with which he can compare or predict his own sales.

A composite *price* index is constructed by this method in the same way as a quantity index. Table 18-4 illustrates the computation of a consumer price index for three types of meat in 1957-1959 (the base period) and the three months ending January 1966, using the price data in Table 18-5. Round steak is chosen as typical of all beef and veal

Table 18-4

CONSTRUCTION OF COMPOSITE INDEX FOR THREE RETAIL MEAT PRICES
BY AVERAGE OF RELATIVES METHOD
(1957-1959 = 100)

Period (1)	Simple Index (1957-1959 = 100)			Weighted Index			Composite Index
	Round Steak (2)	Ham (3)	Frying Chicken (4)	Steak (Column 2×0.57) (5)	Ham (Column 3×0.28) (6)	Chicken (Column 4×0.15) (7)	(Total, Columns 5-7) (8)
1957-1959 Average	100	100	100	57	28	15	100
November 1965	108	109	87	62	30	13	105
December 1965	108	123	84	62	34	13	109
January 1966	107	130	87	61	36	13	110

SOURCE OF PRICE DATA: U.S. Bureau of Labor Statistics, *Estimated Retail Food Prices by Cities*.

prices in its price behavior, while ham represents pork products and frying chicken represents poultry prices. The individual commodity price is then weighted in accordance with the importance of the whole commodity group it represents, rather than by its own individual importance. Of course, actual indexes involve hundreds of commodities and many dates. The steps are similar to those cited above:

1. Divide each price series by its price in the base period (1957–1959 average) to express it as a simple index (Table 18–4, columns 2 to 4).
2. Measure the relative importance of each commodity group in dollars for some normal period. The relative weights in the heading of columns 5 to 7 are based on a hypothetical consumer survey which showed that for every dollar the typical family spent on meat, 57 cents went for beef and veal, 28 cents for pork products, and 15 cents for poultry. The weights preferably apply to the base period, but this is not always feasible. Thus, the U.S. Bureau of Labor Statistics reports its Consumer Price Index with the base 1957–1959 = 100, but since January 1964 it has obtained its weights from a survey of consumer spending patterns made in 1960–1961. (Note that *dollar values*, rather than prices or quantities, are used as weights in the weighted average of relatives method for computing either price or quantity indexes. Also, the weight must be held constant over a period of years; otherwise changes in the weight would affect the level of the index itself.)
3. Multiply the simple indexes (columns 2 to 4) by the weights to obtain the weighted indexes (columns 5 to 7).
4. Add the weighted indexes for each period to get the composite index (column 8). (If the weights are not adjusted to total 1, the last column must be divided by its base-period value to adjust this value to 100.)

Aggregative Method. The aggregative method is more direct than the average of relatives method in bypassing the calculation of simple indexes. Table 18–5 illustrates the construction of a *price* index by the aggregative method. The steps are

1. Choose as weights the physical *quantities* of each commodity produced or consumed in a typical period. In this case, it is the quantity of each of three food items consumed by an average family in a week: 5 pounds of beef and veal, 4 pounds of pork products, and 3 pounds of poultry.

Table 18-5

CONSTRUCTION OF COMPOSITE INDEX
FOR THREE RETAIL MEAT PRICES
BY AGGREGATIVE METHOD
(1957-1959 = 100)

PERIOD (1)	PRICE PER POUND, DOLLARS			COST OF WEEK'S SUPPLY, DOLLARS				COMPOSITE INDEX (Col. 8 ÷ Col. 9.01) (9)
	Round Steak (2)	Ham (3)	Frying Chicken (4)	Steak (Col. 2 × Col. 5) (5)	Ham (Col. 3 × Col. 4) (6)	Chicken (Col. 4 × Col. 3) (7)	Total (Cols. 5-7) (8)	
1957-1959 Average	1.02	0.64	0.45	5.10	2.56	1.35	9.01	100
November 1965	1.10	0.70	0.39	5.50	2.80	1.17	9.47	105
December 1965	1.10	0.79	0.38	5.50	3.16	1.14	9.80	109
January 1966	1.09	0.83	0.39	5.45	3.32	1.17	9.94	110

SOURCE OF PRICE DATA: U.S. Bureau of Labor Statistics, *Estimated Retail Food Prices by Cities*.

2. Multiply each price (columns 2 to 4) by its weight to obtain the weighted prices (columns 5 to 7). The product of price times quantity gives the total cost of each commodity in the "market basket" as its price changes from time to time.
3. Total these products (column 8) to get the cost of the whole market basket.
4. Select a base period (1957-1959 average) and divide the totals by the total in the base period (\$9.01). The results (column 9) are aggregative index numbers. Here they indicate that in January 1966 the combined cost of the three commodity groups was about 110 percent of what it was in 1957-1959.

As a more realistic sample of the aggregative method, Standard and Poor's constructs its price index of 500 stocks by multiplying the current market price of each stock by the number of shares outstanding in the base period (modified by later capitalization changes). This weighted price, or aggregate market value of the original shares, is then totaled for all 500 stocks, and the grand total is divided by the aggregate market value in the base period to obtain the index.⁶

Quantity indexes are computed by the aggregative method in the same way as price indexes, except that quantity and price are interchanged. The varying quantities produced or consumed each month are multiplied by a fixed price in the base year or some other typical period. Hence, only changes in physical volume affect the movements of the index, and the fixed price serves to give each commodity its appropriate importance. Then the sum of the weighted quantities each month is divided by the sum in the average month of the base year to yield the weighted aggregative quantity index.

⁶ The base is set at 1941-1943 = 10 in order to make the current index approximate the average price of all stocks listed on the New York Stock Exchange.

Dollar-value indexes (e.g., department store sales) reflect the movements of both price and quantity, so neither one need be held constant. Furthermore, the original data are already available in the form of dollar values. In the aggregative method, the estimated values for each component of the index are simply added each year. The totals themselves may then be reported, as in gross national product estimates, or they may be divided by a base year value and reported as index numbers, as in the U.S. Bureau of Labor Statistics Index of Manufacturing Production-Worker Payrolls.

The average of relatives method is used when the components are not comparable, as in bank debits and department store sales used in regional business indexes. Here the components are expressed as relatives and then multiplied by arbitrary weights to arrive at the final value indexes.

Formulas for Computing Composite Indexes

The two basic methods of computing weighted index numbers can be expressed in formulas using the following symbols:

For an individual commodity—

- p_0 = price in the base period (e.g., 1957–1959 average),
- p_n = price in current year of the series (e.g., 1967, 1968, etc.),
- q_0 = quantity in the base period,
- q_n = quantity in current year of the series,
- $\Sigma(p_n q_0)$ = sum of (price of first commodity in current year times base-period quantity) plus (price of second commodity in current year times base-year quantity), etc.

The formulas are:⁷

	<i>Average of Relatives Method</i>	<i>Aggregative Method</i>
Price index.....	$\frac{\Sigma(p_n/p_0)(p_0 q_0)}{\Sigma(p_0 q_0)}$	$\frac{\Sigma(p_n q_0)}{\Sigma(p_0 q_0)}$
Quantity index.....	$\frac{\Sigma(q_n/q_0)(p_0 q_0)}{\Sigma(p_0 q_0)}$	$\frac{\Sigma(p_0 q_n)}{\Sigma(p_0 q_0)}$
Value index.....	$\frac{\Sigma(p_n q_n/p_0 q_0)(p_0 q_0)}{\Sigma(p_0 q_0)}$	$\frac{\Sigma(p_n q_n)}{\Sigma(p_0 q_0)}$

The two formulas in each row are identical when the base-period price, quantity, or value is used as weight. That is, multiplying prices by base-year quantities gives the same algebraic result as multiplying price

⁷ These formulas, which use base-year weights, are variants of "Laspeyres' formula," as opposed to "Paasche's formula," which uses current-year weights, or Irving Fisher's "ideal" index, which is the geometric mean of the two.

relatives by the same year's value, etc. If some other period is used as weight, as is often the case, the results will differ somewhat. Thus, the principal U.S. government indexes all use the same 1957–1959 *base* for comparability, while the *weights* for the Consumer Price Index were determined from a survey of consumer expenditures in 1960–1961, the weights for the Wholesale Price Index represent sales of commodities reported in the 1958 censuses, and the weights of the Federal Reserve Board Index of Industrial Production depend on the “value added” by the industry in 1957.

Formulas for quantity indexes are the same as for price indexes with p and q interchanged.

Comparison of Average of Relatives and Aggregative Methods

The average of relatives and aggregative methods often yield identical results, as described above. Then which is the better one to use?

The aggregative method is the simpler and the more easily understandable of the two, so it may be used whenever appropriate weights (i.e., quantities for a price index) are available and when only the composite index is needed.

The average of relatives method, on the other hand, must be used when:

1. It is desired to compare the individual components in the form of relatives, as in *The New York Times* Index of Business Activity. The first step in this method produces these relatives directly.
2. The available weights are in value form, as in the Federal Reserve Board index, which applies the “value added by manufacture” for a group of related items as a weight for the production of a single representative item. It is usually easier to obtain dollar values as weights than it is to find quantities or composite prices.
3. The component series are already in the form of relatives, as in combining several segments of the Federal Reserve Board Monthly Index of Industrial Production for comparison with a particular industry.

Since one or more of these conditions usually exist, the average of relatives method is more widely used than the aggregative method.

TESTS OF A GOOD INDEX NUMBER

A businessman must often refer to index numbers in gauging the state of the economy and in making necessary day-to-day decisions for

the control and planning of his operations. Yet he cannot accept an index uncritically at its face value without inquiring into its characteristics and limitations. Appearances are deceiving, and the official names of indexes are often little more than general guides to their nature.

If one makes any regular use of an index, therefore, it is surely worthwhile to write the publisher for a description, or at least to check one of the publications at the end of this chapter that provide a critical analysis of the major indexes. One should also appraise the reliability and reputation of the compiler. For example, the leading federal statistical agencies have improved their indexes tremendously, while on the other hand, certain regional chambers of commerce publish extremely crude indexes of business activity in their areas.

In studying the nature of an index it is particularly important to apply the following tests, which determine whether the index is suitable for your need: (1) the purpose of the index, (2) selection of the sample, (3) choice of the base period, (4) selection of weights, and (5) statistical adjustments.

Purpose of the Index

The exact purpose that an index number is intended to serve should be clearly understood by the reader. Thus, the Consumer Price Index is intended to measure the cost of a fixed bill of goods and services purchased by lower-income urban workers; it does *not* claim to measure the cost of living of consumers generally, as is often misconstrued. Again, The Dow-Jones Averages purport to measure the relative price changes of "blue-chip" market leaders, not the stock market generally. In similar fashion, the F. W. Dodge Corp. index of construction contracts awarded was developed to indicate relative changes in the value of contract building. It cannot be used to measure changes in the physical volume of construction nor changes in the value of construction put in place.

If a single index number proves inadequate, the use of several related indexes may fulfill a given need. For example, in analyzing monthly changes in regional business activity, it is useful to supplement a composite business index with indexes of employment, payrolls, construction contracts, retail sales, and the like that reflect changes in component elements of business.

Selection of the Sample

The second test of a good index number arises from the statistical requirement that the data must provide a representative sample, unless,

of course, they cover the entire field. The principles for selecting a sample have been treated in Chapter 14. It is of the utmost importance that the data collected for constructing index numbers conform to these principles. Otherwise, no valid generalizations can be drawn from the results.

The following sampling plan is an effective and appropriate one in selecting a sample of items to include in an index number.

First, divide the commodities into a large number of small groups or strata. Each group should comprise a closely related line of products that might be expected to move fairly uniformly in price, quantity, or value, as the case may be. Weights must be available for these groups. This stratification permits accurate weighting and flexible grouping into main categories as desired.

Then select from these groups a typical list of items to include not only all of the most important articles but also some that are typical of every category of goods in the group both in physical characteristics and price behavior in the case of a price index. Of course, each item must be precisely identified. The prices are then weighted and the products totaled to form group indexes, and the latter are again combined to provide the overall index. The result may be called a highly *stratified judgment sample*.

In groups or parts of groups where there is little basis for selection, as when there are many items of minor or relatively equal importance, each tenth, twentieth, or some other numbered item may be taken from the list.⁸ This is a *systematic*, rather than a judgment, sample.

In any case, the proper selection of a typical cross section of items is the most crucial step in the entire process. Many regional "general business" indexes and others fail in this respect—they just do not measure what they purport to represent.

The number of items selected in each group may vary from one to twenty or more, depending on the group's importance and diversification. For all groups combined, several hundred items should be priced to constitute a sample of adequate size. The Bureau of Labor Statistics, for example, includes about 400 items in its Consumer Price Index,⁹ while the Standard and Poor's index includes the prices of 500 common

⁸ Alternatively, the items may be selected with "probability proportional to size," size being defined as the relative weight of the item. See M. Wilkerson, *Sampling Aspects of the Revised CPI* (Washington, D.C.: U.S. Bureau of Labor Statistics, October 1, 1964), p. 12.

⁹ On the other hand, some 2,200 items are included in the Bureau's Wholesale Price Index in order to insure the reliability of its many component indexes.

stocks. A smaller number might be used, however, for items that are fairly homogeneous as to type and price behavior.

Choice of a Base Period

The base of an index showing changes from time to time may be any period that provides the most suitable standard for comparison. There are a number of criteria for the selection of such a base. The most important of these are (1) normality of the period, (2) trustworthiness of the data in the period, (3) comparability with existing index numbers, and (4) inclusion of census years for bench-mark data.

Normality of Period. It is frequently held that the base period should be one that is "normal" or "average"; that is, a period when the level of the data is about midway between the peaks and troughs of business cycles in that era. A period of very high prices, for instance, should not be used as the base because the influence of the most inflated components would be disproportionately low in other periods. In contrast, if a period of very low prices were used as the base, the influence of the most depressed components would be disproportionately high in other periods. Thus, neither the depression years 1931–1934 nor the war years 1942–1945 or 1950–1953 are as suitable base periods as are the more average levels of 1935–1939, 1947–1949, or 1957–1959. These three- to five-year periods have been chosen for U.S. government indexes in preference to a one-year base because the longer periods tend to iron out the year-to-year irregularities.

Trustworthiness of Data. Source materials have become generally more accurate and comprehensive in recent years, so that a recent period is more likely to provide a reliable base than an earlier period. The Bureau of Labor Statistics Wholesale Price and Consumer Price Indexes and the Federal Reserve Board Index of Industrial Production, for example, have all been revised in recent years to include new products and to embody new weights reflecting changed production and consumption patterns. At the same time the older base periods were replaced by a 1957–1959 base, which more nearly encompasses both the recently developed products and the particular years for which the weights are computed.

Comparability with Other Index Numbers. The base for a new index number is often chosen to coincide with that of existing index numbers with which the new one is most likely to be compared. Index numbers are not directly comparable unless their base periods are identical. For this reason the Office of Statistical Standards in the Bureau of the Budget has endeavored to standardize governmental indexes on a

1935–1939, 1947–1949, and 1957–1959 base in these successive decades.

Inclusion of Census Years. Since it is preferable to use base-year weights as nearly as possible,¹⁰ the base period should include census years for which bench-mark data are available as weights. The base period 1957–1959, for example, includes the 1958 Census of Manufactures, the 1958 Census of Business, and the 1959 Census of Agriculture.

Weights

Earlier in this chapter, weights were defined and used in calculating composite index numbers. Here the problems of selection of weights, type of weights, shifting weights, and weight bias are discussed.

Selection of Weights. Weights may be selected to represent either the importance of a specific commodity or the importance of the entire economic group of which it is typical. In the latter case, one might include in a production index of house furnishings the relative for a standard type of domestic wool rug weighted by the total value of all sorts of similar rugs rather than to include a large number of different rugs and weight each one according to its own specific importance. This group weighting system is used in the Federal Reserve Board Index of Industrial Production and the Bureau of Labor Statistics Consumer Price Index, as described later in this chapter.

Weights should also be appropriate to the purpose of an index. An average of relatives price index for a company's inventory, for example, should be weighted by inventory values; a price index of goods sold should be weighted by sales values; while a consumer price index should be weighted by consumer expenditures.¹¹

Physical Quantities or Values as Weights. The factors used as weights for a given index number depend upon the method of construction and the kinds of data being employed. If it is an index number of prices and the aggregative method is used, that is, a method which adds the actual weighted prices, the weights must be *quantity* data of some kind, never value. Value includes the effect of price, since it equals price times quantity. Its use as a weight in an aggregative index would actually have the effect of squaring the prices, which would give undue

¹⁰ U.S. Bureau of the Budget, Division of Statistical Standards, *Recommendations on Postwar Base Period for Index Numbers* (March 14, 1951), p. 2.

¹¹ Weights may be rounded off to two or three significant figures, or even one figure for minor items, since an appreciable difference in weights will affect an index but little.

importance to changes in the larger prices. Conversely, an aggregative quantity index would be weighted by *prices*. For an average of either price or quantity relatives, on the other hand, *value* weights should be used, as illustrated in Table 18-4.

Whether the weights used will be quantities or values may, however, depend upon the availability of data. For most kinds of commodities, exchange values in dollars are more likely to be available than quantities. Values must also be used for group weights, where the items are in different units. In these cases, the weighted average of relatives method should be used.

Constant or Variable Weights. Index numbers are designed to show changes only in the variable being measured—a price index, for instance, should isolate changes in price from changes which may be due to quality changes and other factors. None of the factors in the computation except prices should be allowed to fluctuate. The weights, therefore, should usually be kept constant for an extended period. If prices and weights were allowed to vary simultaneously, the resulting index numbers would reflect changes due to both factors, and no one could tell what part of the final result was due to variations in prices and what part was due to variations in the weights.

This raises the question: If the weights are to be held constant for extended periods, which specific period should they represent? In the examples used as illustrations of method, the weights were quantities or values in the period used as the base of the index numbers, but this is not necessarily the best procedure to follow in every case.

The importance of commodities may change during relatively short periods so that, if weights of an early period are used, there is a danger that the current index number will not accurately reflect the present relative importance of its several constituents. For instance, the cost of purchasing and maintaining a color television set is an important element in present-day cost of living that did not exist a few years ago.

When it is definitely known that the constituents of the index are changing in importance, weights should be revised from time to time. Too frequent revisions, however, tend to impair the usefulness of an index number, so that ordinarily no change should be made as long as the weights are approximately correct. In long-established indexes the weights have been changed at intervals of about ten years.

Bias Due to Weighting. Bias due to methods of weighting is almost certain to occur in some degree. In this sense "bias" means that the index number tends to understate or overstate the degree of change because of the failure of the weights to represent accurately the relative

importance of shifts in the items included. Price indexes are generally based on the cost of a fixed bill of goods, but people actually buy different quantities as prices change. The probable bias of any index due to shifts in consumption patterns and the like should be carefully considered before it is used in a major policy decision.

Statistical Adjustments

Most composite monthly indexes should be adjusted statistically to show the cycles and the long-term trend in the underlying data and to eliminate seasonal and irregular movements. (These adjustments will be discussed in Chapters 20 and 21.) That is, (1) the data should be adjusted for seasonal and calendar variations if necessary; (2) the resulting figures should be smoothed by moving averages (see "Months for Cyclical Dominance" in Chapter 21), so that the series will show more consistent trend-cycle changes from month to month than meaningless zigzag irregularities; and (3) a dollar value series should be deflated by a price index if it is desired to show physical volume changes (Chapter 19). It is also desirable to determine whether the index is typically a leading, coincident, or lagging indicator at business cycle turning points. (See U.S. Department of Commerce, *Business Cycle Developments*, monthly.)

Monthly business indexes should also be checked against more complete annual data or quinquennial censuses of manufactures and other censuses in order to adjust the general trend of the monthly series to these more accurate bench marks. Otherwise, a monthly index based on sample data will develop a cumulative upward or downward bias over the years which will destroy its validity for long-term comparisons.

REVISIONS OF INDEX NUMBERS

Substitution of Items

Changes in production, distribution, habits of consumption, and a variety of other economic factors sometimes necessitate substitutions in the items included in an index, in its list of respondents, or in the specifications of the items included. For example, the changeover from oil to gas heating led the Bureau of Labor Statistics in 1958 to substitute a 30-gallon domestic hot-water gas heater for a similar oil heater in computing its Wholesale Price Index. The availability of new and better data may also make it desirable to revise established index numbers, as described above. When interpreting the movement of index numbers it

is essential that these changes be kept in mind, for the particular method of revision may make a great deal of difference in the final result.

Changing the Base Period

The base period of an index number may need to be changed in either of the following situations: (1) When index numbers based on different periods are to be compared, it is necessary to shift one index to the same base period as the other, so that changes in the two will be measured from the same point in time. (2) It may be desired to shift the base of a series to some arbitrary reference date such as 1960 in order to compare subsequent changes with conditions at that time.

A series can be shifted to a new base by multiplying each of its index

Table 18-6
SHIFTING THE BASE OF PRICES PAID BY FARMERS
FROM 1910-1914 TO 1957-1959 FOR COMPARISON
WITH THE CONSUMER PRICE INDEX

	PRICES PAID BY FARMERS FOR FAMILY LIVING ITEMS		CONSUMER PRICE INDEX
	1910-1914 = 100 (1)	1957-1959 = 100* (2)	1957-1959 = 100 (3)
1957	282	99	98
1958	287	100	101
1959	288	101	101
⋮	⋮	⋮	⋮
1964	300	105	108
1965	305	107	110

* Obtained by multiplying column 1 by $100/285.7$ to shift the 285.7 value for the 1957-1959 average to the 100 level.

SOURCE: *Survey of Current Business*.

numbers by $100/X$, where X is the index number for the period selected as the new base. That is, $X \cdot 100/X = 100$. Since each of the indexes is multiplied by the same constant factor, the *relative* fluctuations of the series remain unchanged.

To illustrate, in Table 18-6 the base period for prices paid by farmers for family living items has been shifted from 1910-1914 to 1957-1959 for comparison with changes in the Consumer Price Index since that period. Since the original index of prices paid by farmers averaged 285.7 in 1957-1959, the whole series has been multiplied by $100/285.7 = .3500$ to shift the 1957-1959 average to 100 (column 2), the same as for the Consumer Price Index. Note that index numbers

for the base years must average 100. The last two columns show that from the 1957–1959 average to 1965, prices paid by farmers advanced only 7 percent as compared with 10 percent for consumer prices generally, even though the original farm price index increased by more points than the Consumer Price Index.

Splicing Two Series

It is often necessary to splice two series to form a continuous series, as when the specifications of a commodity in a price index are changed. Any two series may be spliced, provided they are both available for the same year. For example, the BLS Wholesale Price Index might be said to include everything but the kitchen sink. This is not true. It includes an enameled steel sink, but the price of a new reporting company was added to its sample in November 1958. As a result, the typical price had to be shifted from \$13.39 (or an index of 100.8 on the 1957–1959 base) to \$13.13 in that month. Table 18–7 shows how to continue the original price index (column 2) for the sink by splicing the new price (column 3) onto it. The new price of \$13.13 in the overlapping month November 1958 must be shifted not to 100 but to 100.8, the index for that month. The new price series, therefore, is multiplied by $100.8/\$13.13$, as shown in column 4. The spliced series in column 5 (combining columns 2 and 4) now shows enameled steel sink prices continuously throughout the period, although the actual sample price shifts in November 1958.

As another example, the new car component of the Consumer Price Index (based on a standard-size Chevrolet, Ford, and Plymouth) became outmoded in 1960 with the widespread introduction of compact cars, whose price behavior differed from that of standard-sized models. Hence, the Bureau of Labor Statistics introduced the prices of four small cars (Rambler, Falcon, Valiant, and Corvair), linking the new series onto the old in October 1960 so that level of the index was not affected by the lower price of the compact cars.¹²

Strictly speaking, an index which is being shifted to a new base should be composed of the same items during the whole period of the index. Yet the most common use of base shifting is to link a current index containing one group of items to an earlier-period index containing a similar but not identical group of items. This procedure is legiti-

¹² O. A. Larsgaard and L. J. Mack, "Compact Cars in the Consumer Price Index," *Monthly Labor Review* (May 1961).

Table 18-7

SPlicing TWO PRICE SERIES
REPRESENTING AN ENAMELED STEEL SINK

(Prices in Dollars; Indexes on 1957-1959 Base)

	ORIGINAL SAMPLE OF REPORTING COMPANIES		ENLARGED SAMPLE OF REPORTING COMPANIES		SPliced SERIES
	Price (1)	Index (2)	Price (3)	Index (4)	Index (5)
September 1958	\$13.194	99.4			99.4
November 1958	\$13.39	100.8	\$13.13	100.8	100.8
June 1959			\$12.71	97.6	97.6

SOURCE: U.S. Department of Labor, *Wholesale Prices and Price Indexes, 1958*, Bulletin No. 1257, (July 1959), pp. 225 and 230 (item #1053-11), shifted to 1957-1959 base.

mate if the old and new groups of items may be considered to be representative of the same population. This is true of the above example. In case the components of an index have changed more radically from time to time, however, as in the Cleveland Trust Company Index of Industrial Production from 1790 to date, the index loses its homogeneous character.

SOME IMPORTANT INDEXES

There are many more business indexes in common use than can be treated here. Hundreds of these are described in the readings at the end of the chapter. We will discuss only three major indexes—their construction, uses, and limitations—to illustrate the typical problems involved. These are the consumer and wholesale price indexes of the U.S. Bureau of Labor Statistics and the industrial production index of the Federal Reserve Board. The base period for all these indexes is 1957-1959 = 100.

Consumer Price Index

"The Consumer Price Index is a statistical measure of changes in prices of goods and services bought by urban wage earners and clerical workers, including families and single persons."¹³

¹³ This is the definition of the "new series," first published in January 1964. See U.S. Department of Labor, *The Consumer Price Index (Revised January 1964)*, *A Short Description* (September 1964) and *Monthly Labor Review* (August 1964), p. 967 f, for further details.

The index is computed by the weighted average of relatives method¹⁴ using constant weights. Prices are measured monthly or quarterly, and the aggregate cost of a fixed bill of goods and services is compared with that in the base period 1957–1959. Since the quantities represent not only consumption of the 400 goods and services actually priced but also consumption of related items for which prices are not obtained, the total cost of the “market basket” represents a broad sector of total consumer spending for goods and services.

The prices collected for this index are retail prices charged to consumers for “food, clothing, automobiles, homes, housefurnishings, household supplies, fuel, drugs, and recreation goods; fees to doctors, lawyers, beauty shops; rent, repair costs, transportation fares, public utility rates, etc.” These prices include sales and excise taxes as well as real property taxes but not income or personal property taxes.

The 400 goods and services comprising the “market basket” of items sampled are representative of the typical goods and services purchased by urban wage and clerical worker families and single individuals living in urban areas with a 1960 population of 2,500 or more persons. These families and single workers comprised about 56 percent of the people living in urban places and about 40 percent of the total U.S. population in 1960. The index is designed to measure *only* changes in prices of the same “market basket” through time, *not* to measure changes in the composition of different “market baskets” or changes in consumers’ standards of living.

Periodically, the bureau conducts Consumer Expenditure Surveys to determine the pattern of expenditures for goods and services by wage earners and clerical workers. The last survey was conducted for the years 1960–1961 in 66 urban areas, which were chosen to represent all urban places in the 50 states. From the data collected, the bureau revised the quantity weights used to compute the “new series” index and objectively select the 400 items to be included.

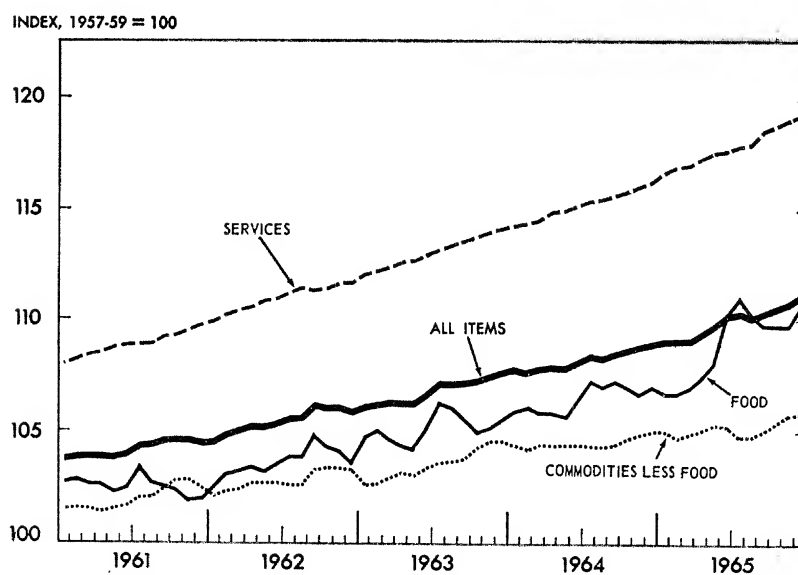
All items purchased by wage earners and clerical workers were grouped or stratified into “expenditure classes.” The items included in each of the 52 expenditure classes, which define the sampling strata, were primarily determined by grouping items which in a general way serve the same human needs. Items were selected with probability proportional to their relative importance as compared with total expend-

¹⁴ Three variants of this method are actually used: (1) the “average of price relatives for reporting outlets,” (2) the “relative of average prices for identical outlets,” and (3) the “relative of average prices for all reporting outlets.” See M. Wilkerson, *Sampling Aspects of the Revised CPI* (U.S. Bureau of Labor Statistics, October 1, 1964).

itures for all items. According to this plan, the most important items were certain to be selected as their relative importance was greater than the selecting interval.

The urban places in which the bureau collects price data for the CPI also were selected by probability sampling. The primary sampling units were 50 standard metropolitan statistical areas. These units were stratified by broad region and by size of population into 12 strata. The 12 largest areas were selected with "certainty," again because their size

Chart 18-1
CONSUMER PRICES



SOURCE: Department of Labor.

exceeded the selecting interval. Six large additional metropolitan areas were added in January 1966.

The relative importance of each area in the CPI is determined by the proportion of total wage-earner and clerical-worker population it represents to the total for all areas represented in the CPI, based on 1960 Census data. Chart 18-1 shows the changes in the index and in three major components for 1961 to 1965.

Uses of the Consumer Price Index. The original Cost of Living Index was established at the close of World War I to aid in the adjustment of ship-builders' wage rates. Since that time the index has become an increasingly important aid to unions and management in adjusting wages to take account of changes in consumer prices.

The most important subsequent impetus to the use of the index for this purpose was its designation as a basis of wage-rate escalation in the contract signed by the United Automobile Workers and the General Motors Corporation in May 1948. Since then the agreement has been extended several times and is now due to expire in September 1967. The escalator clause now provides for a 1 cent an hour quarterly wage adjustment for each 0.4 point change in the CPI. Other collective bargaining agreements follow a similar pattern. For example, the agreement between the retail food industry in Los Angeles and the Building Service Employees' International Union, to expire May 1969, provides for a 1 cent an hour quarterly wage adjustment for each 0.5 point change in the CPI.¹⁵

After each of these major agreements, many other contracts were signed on the same basis, frequently without any examination of the reasonableness of the relationship of wage-rate changes to index changes in each particular situation, or without full realization of the effects of arbitrarily accepting a ratio based on some other firm's or union's experience. Whatever the type of escalator employed, however, it is important to both sides in a bargaining group that the procedure be adjusted to each particular situation.

Escalator clauses based on the CPI are used not only to adjust wage payments but also to adjust rents, pensions, alimony, fiduciary payments, and many other types of contracts. Finally, the CPI is widely cited as an indicator of inflation as it affects the consumer. It serves, therefore, to measure the purchasing power of the consumer's dollar.

The Consumer Price Index also has limitations which should be carefully considered: (1) It measures changes only in a fixed bill of goods and services, but not changes in the standard or manner of living. (2) It does not always reflect gains due to the improvement in the quality of manufactured products. Hence, it is claimed to overstate the true rate of inflation.¹⁶ Conversely, in wartime conditions of material shortages, it fails to reflect the full inflationary effect of black-market prices, quality deterioration, and substitution of more expensive grades for cheaper grades of products. (3) While it measures changes in consumer prices from time to time, it cannot be used to compare prices between different places at a single point in time. Geographic differences may be measured by comparing the individual prices compiled for

¹⁵ See *Major Collective Bargaining Agreements: Deferred Wage Increase and Escalator Clauses*, U.S. Department of Labor Bulletin No. 1425-4 (January 1966).

¹⁶ See W. Allen Wallis, *Journal of the American Statistical Association* (March 1966), pp. 1-10; also, *Monthly Labor Review* (September and November 1961), articles by Milton Gilbert and Ethel Hoover, respectively.

the Consumer Price Index, but not the index itself. (4) The index measures changes in prices only for the worker group in urban areas. It should not be used without modification for other income groups or for families living in nonurban areas. (5) Since the index represents an average family's consumption pattern, it may not represent the experience of any specific family or individual.

Wholesale Price Index

The Wholesale Price Index of the Bureau of Labor Statistics measures the average rate and direction of movements in commodity prices at primary-market levels—that is, at the point of the first commercial transaction for each commodity—and specific price changes for individual commodities and groups of commodities.¹⁷ The prices used in the index are those representing all sales of goods by or to manufacturers or producers, or those in effect on organized commodity exchanges. Therefore, it represents producers' prices or primary-market prices rather than those charged by wholesalers.

Prices for approximately 2,200 separate specifications of commodities are included in the index. To obtain "real" or "pure" price changes not influenced by changes in quality, identical lists of commodities defined by precise specifications are priced from month to month. Prices are adjusted for trade and quantity discounts, as well as cash and seasonal discounts when these are customary. Excise taxes are excluded. These prices are obtained from some 2,000 companies which are asked to quote the prices they actually charge for a specific commodity to a given type of buyer on a particular day, usually the Tuesday of the week including the fifteenth of the month. Some quotations from trade journals and market reports are also used.

Because the commodity population is so large, the index is based on a sample of commodities, a sample of specifications for the commodities, and a sample of reporting sources. The individual items are selected as the most important in each field and as those believed to represent the price movements of other closely related commodities. The sample is thus a highly stratified, selected group, rather than a random sample. The broad coverage of 2,200 items permits the development of reliable subindexes for many small subdivisions of the economy.

The index is calculated fundamentally as a weighted average of price relatives in which the weights are based on net sales values of commodities reported by the Census of Manufactures, the Census of Mineral

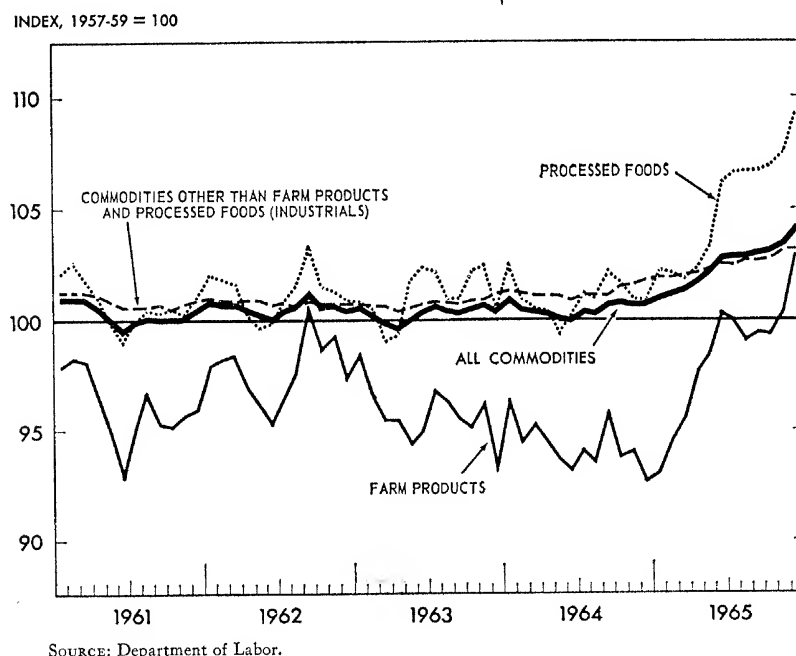
¹⁷ See U.S. Department of Labor, *Wholesale Prices and Price Indexes, 1962*, Bulletin No. 1411 (June 1965), pp. 7–15.

Industries, and other sources for 1958. Each item has a weight which includes its own weight based on its sales in 1958 and the weight of the other items it represents in the index. The weights will be revised to reflect the relative importance of commodities in later censuses scheduled at five-year intervals.

The overall index is divided into the broad categories of industrial commodities, and farm and food products, as shown in Chart 18-2. Special wholesale price indexes are reported by stage of processing and by durability of product. In addition, separate indexes are published each month for 15 major groups, 86 subgroups, about 250 product classes, and for most of the individual series.

The Bureau of Labor Statistics also prepares a Weekly Wholesale

Chart 18-2
WHOLESALE PRICES



Price Index based on actual weekly prices of a sample of about 200 of the commodities included in the monthly index and on estimates of the prices of the other commodities. This index may be used to give interim estimates of the monthly index.

Uses of the Wholesale Price Index. The Wholesale Price Index is one of the basic business barometers used to measure the economic health of the nation. It is also used as a price deflator or as a

purchasing-power index, reflecting changes in the value of the dollar. The important application of price indexes in deflating value series is described in Chapter 19.

This index, or any of its component indexes, may be used for comparison with series of individual business data. For example, the General Electric Company provides its purchasing offices with a price index of commodities purchased by the company, weighted by their importance to the company, and compares this with the BLS wholesale price index for industrial commodities.¹⁸

One of the most frequent uses of the Wholesale Price Index is as an escalator—that is, as the basis for adjusting contractual payments or values for changes in the value of the dollar. Long-term production contracts include escalator clauses as guarantees against losses due to increases in the prices of materials and other costs. Rentals on long-term leases are also often adjusted by this index.¹⁹

There are limitations to the wholesale price indexes which must be kept in mind when using them: (1) They measure primary-market prices, not wholesalers' prices as the name implies. (2) Most of the indexes relate to national coverage and hence should be used with caution in interpreting local or regional data. (3) Since they relate to changes of a given specification, they cannot be used with retail price indexes to calculate margins. (4) The indexes do not include any of the services, such as rent, transportation, or communications.

Industrial Production Index

The Federal Reserve Board's Monthly Index of Industrial Production is one of the most widely used of the country's economic indicators. It measures changes in the physical volume of output of factories, mines, and gas and electric utilities from 1919 to date.²⁰

The industrial production index includes 207 series expressed in physical terms—units, tons, yards, board feet, and the like—reflecting the production of American industries or data which represent such series. Where physical output data are lacking, other series which are believed to fluctuate in the same way as output data are substituted. Such series include volume of shipments, production-worker man-hours, materials consumed in production, etc. About 49 percent of the

¹⁸ C. Willard Bryant, "Planning to Meet Materials Shortages," *Purchasing* (August 1953), pp. 81–83.

¹⁹ See "The Use of Price Indexes in Escalator Clauses," *Monthly Labor Review* (August 1963).

²⁰ See "Industrial Production: 1957–59 Base," in *Federal Reserve Bulletin* (October 1962).

weight of the monthly index is represented by man-hour data adjusted for estimated changes in output per man-hour. The balance is based on production and shipments data and miscellaneous measures.

The component series of the index are combined with weights based on value added by the industry in 1957, mainly as shown by the Census Annual Survey of Manufactures. The composite index is calculated as a weighted average of relatives. It is expressed in terms of the 1957–1959 average as a base, for comparability with other index numbers. The index is published for four broad classifications having the following relative importance in 1957–1959: durable manufactures, 48 percent; non-durable manufactures, 39 percent; mining, 8 percent; and utilities, 5 percent. A separate classification is made between the output of consumer goods, output of equipment (including ordnance) for business and government use, and materials. Indexes are also reported for some 25 major industrial groups, following the latest Standard Industrial Classification of the U.S. Bureau of the Budget, and for some 175 subgroups. This great number of industry series permits flexible grouping for most desired comparisons.

The monthly production series are adjusted to levels shown by bench-mark production indexes based on the Censuses of Manufactures and Minerals and for interbench-mark years, mainly Census Annual Surveys. These adjustments are made periodically, and usually during a revision of the index. Between revisions, the levels of the monthly indexes are checked against independently compiled data, such as deflated manufacturers' shipments adjusted for inventory change and electric power used by the manufacturing and mining industries.

Uses of the Industrial Production Index. The major use of the Index of Industrial Production is as an indicator of the economy's output. It is the most sensitive and reliable indicator we have to answer the questions "Is production increasing or decreasing?" and "In which industries are major increases or decreases occurring?" Chart 18–3 shows its movements in describing the pattern of business changes from 1957 to 1966. The index is widely used in conjunction with other series for both forecasting and guidance in administrative decisions. For example, it is compared with figures on unemployment to obtain estimates of the country's total number of unemployed workers that may be associated with different levels of production. It is also compared with data on inventories and prices.

The detailed industry indexes serve as very useful comparisons or bench marks in studying the production of individual companies. The individual indexes are also useful in comparing growth rates in different sectors of the economy.

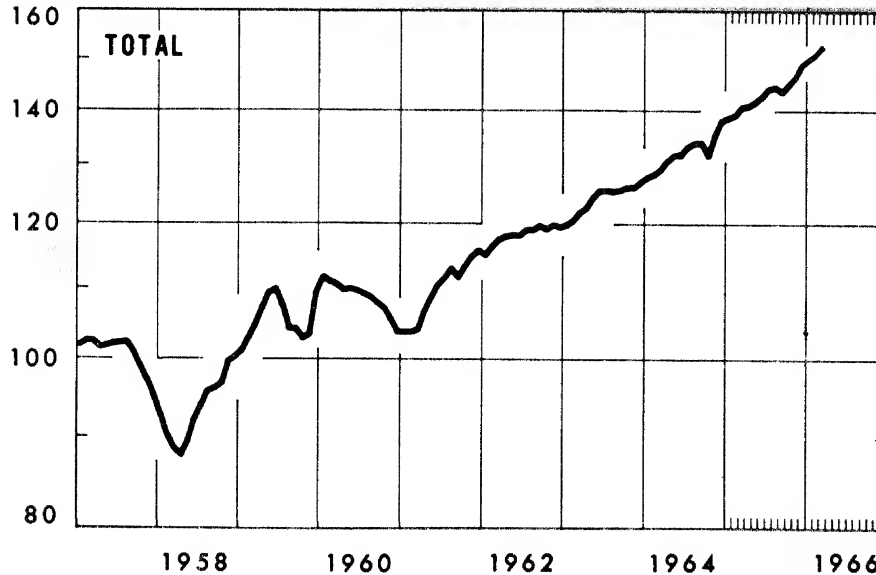
One limitation of the industrial production index is its restriction to manufacturing, mining, and utilities, which keeps it from serving as a measure of total production. Agriculture, construction, transportation, communication, and other services are not included. Another limitation is that changes in man-hours and other indirect measures of industrial activity sometimes do not reflect accurately the changes in physical volume of production, particularly in times of war and postwar reconversion.

Chart 18-3

INDUSTRIAL PRODUCTION

RATIO SCALE
1957-59=100

MONTHLY, SEASONALLY ADJUSTED



SOURCE: Federal Reserve Chart Book, April 1966.

SUMMARY

Index numbers express the changes in a variable relative to some base taken as 100. They are particularly useful in comparing different series and in combining a group of series in a single summary figure. Most indexes are designed to show changes in price, quantity, or value (price times quantity), either from time to time or from place to place.

A simple index or relative is constructed by dividing a single series by its base figure and multiplying by 100.

Composite indexes should ordinarily be weighted arithmetic means of their components. A composite price or quantity index may be constructed by two methods: (1) In the weighted average of relatives

method, the relatives are first computed for each series as described above and then multiplied by value weights expressed as decimal fractions of the total weight. The sum of the weighted relatives is the composite index. (2) In the aggregative method, the changing prices are multiplied by fixed quantity weights (or vice versa for a quantity index). The resulting products are then totaled, divided by the product in the base period or place, and multiplied by 100. The weights usually represent the importance of a component in the base years or some other normal period. In a *value* index the dollar values of each component are simply added in the aggregative method or else the components are expressed as relatives and multiplied by arbitrary weights before being totaled.

The aggregative method is the simpler of the two, but the average of relatives method is preferable when individual series are to be compared, when available weights are in value form, or when the component series are expressed as relatives.

The following tests of a good index should be applied in appraising the validity of an index for some specific use: (1) The purpose of the index should be clearly defined. (2) The items included must be specifically related to the purpose and must be a representative sample of the population being measured. (3) The base period should be a fairly normal one, adequate in length, easy to recall, and one used by comparable indexes. Trustworthy data and census bench marks should be available for this period. (4) Appropriate quantity weights should be used in an aggregative price index, and vice versa, or value weights in an average of relatives index. Weights must be held constant, but should be revised every decade or so as the importance of the components changes appreciably. The probable bias due to weighting should also be considered.

Items may be substituted for others in an index, as necessary, by proper "linking." An index number may be changed to a new base or spliced onto a similar series by multiplying or dividing by a constant factor without changing the relative movements of the index in any way.

The construction, uses, and limitations of three major indexes are discussed to illustrate typical examples. The consumer and wholesale price indexes of the Bureau of Labor Statistics represent broad samples of prices at the retail level and the primary market level, respectively. They are widely used as economic indicators, as deflators of value series, and as escalators in contracts. The proper use of the Consumer Price Index in wage contracts is particularly important.

The Federal Reserve Monthly Index of Industrial Production is an important and sensitive measure of general industrial activity. It represents the physical volume of production, shipments, or man-hours in the manufacturing, mining, and utility industries.

Many other indexes are described in the Selected Readings below.

PROBLEMS

1.
 - a) Briefly describe three broad types of index numbers that are used to measure changes in business and economics.
 - b) In your opinion, what is the one most important use of (1) simple index numbers and (2) composite indexes? Give reasons for your choice in each case.
 - c) Cite the principal limitations of index numbers.
2.
 - a) Compute a composite index of grain prices for the data below by the average of relatives method, with 1964 = 100, using base-year weights.
 - b) Compute a composite price index by the aggregative method, using the same base.
 - c) Compare the merits of the two methods in this case.

	Price (Dollars per Bushel)		Production (Billions of Bushels)	
	Wheat	Corn	Wheat	Corn
1964	\$1.92	\$1.23	1.28	3.48
1965	1.70	1.25	1.32	4.08
1966	1.88	1.30	1.31	4.10

NOTE: Price is wholesale, average, all grades; production is crop estimate as of December 1, 1966.

SOURCE: *Survey of Current Business* (February 1967,) pp. S-27 and 28.

3. Using the data in Problem 2 above:
 - a) Compute a composite index of grain *production* by the average of relatives method, with 1964 = 100, using base-year weights.
 - b) Compute a composite production index by the aggregative method, on the same base.
 - c) Compute an index of the *value* of grain production, on the same base.
4. As a purchasing agent for the Steel Products Company, you wish to compile a composite price index for iron and steel purchased, based on the following data:

PURCHASES OF THE STEEL PRODUCTS COMPANY

	Price per Ton			Thousands of Tons Purchased		
	Pig Iron	Steel Scrap	Steel Billers	Pig Iron	Steel Scrap	Steel Billers
1966	\$61	\$54	\$81	10.0	3.0	5.0
1968	66	38	94	11.0	2.1	5.5
1970	66	34	95	10.7	3.6	2.7

NOTE: Pig iron and steel scrap are in long tons; steel billers are in short tons.

- a) Compute a composite index for iron and steel prices each year by the average-of-relatives method with 1966 = 100, using value purchased in 1966 as weights.
 - b) Compute a composite price index by the aggregative method, using the same year for the base and for weights as above.
 - c) How do the indexes obtained in *a* and *b* differ? Explain. What is the chief advantage of each method in this case?
5. a) Compute a composite index of the quantity of iron and steel purchased each year, from the table above, using the average-of-relatives method. Take 1966 as base, and use 1966 values as weights.
 - b) Compute a composite index of the dollar *value* of iron and steel purchased each year, with 1970 = 100.
 - c) Explain the significance of the quantity and value indexes computed above, as opposed to the price index.
6. As a cost analyst with a petroleum company, you are asked to compile an annual index of oil well drilling costs beginning in 1957, with 1957–1959 as a base. You determine that the cost of drilling an oil well is made up of approximately 60 percent labor and 40 percent material, and you decide that the following data adequately represent these elements.

OIL WELL DRILLING COSTS, 1957–1965

Year	Average Hourly Earnings, Petroleum Workers	Wholesale Price Index, Metals and Metal Products (1957–1959 = 100)
	(1)	(2)
1957	\$2.77	99.7
1958	2.84	99.1
1959	2.99	101.2
1960	3.02	101.3
1961	3.16	100.7
1962	3.19	100.0
1963	3.32	100.1
1964	3.37	102.8
1965	3.47	105.7

SOURCE: Survey of Current Business (May 1966) and supplement, *Business Statistics*, 1965.

- a) List the indexes of drilling costs, along with any columns of computations needed.
 - b) What was the percent increase in drilling costs from 1957 to 1965? If 1965 were the base of the drilling cost index, what would the 1957 index be? If labor and materials each made up half of drilling costs, would the index be higher or lower in 1965 than that shown? Why?
 - c) What more refined indexes might you be able to find, to replace those used here, so as to provide a better index of your company's drilling costs?
7. The Bureau of Business Research of the University of Texas published a monthly *Index of Texas Business Activity* with the following description:

"1947-49 average = 100. Components: Retail sales, industrial electric power consumption, miscellaneous freight carloadings, building authorized, crude petroleum production, ordinary life insurance sales, crude oil runs to stills, total electric power consumption (weighted 46.8, 14.6, 10.0, 9.4, 8.1, 4.2, 3.9, and 3.0, respectively, and adjusted seasonally)." Each component was expressed as an index with 1947-1949 = 100 before being weighted. Apply our tests of a good index number to give an appraisal of this index, listing its good and bad points.

8. Index numbers are ordinarily based on samples, so that care must be exercised to insure that the items included in the index are typical of the population.
 - a) Describe the population represented by (i) an index of prices received by farmers, (ii) an index of industrial building costs, (iii) an index of manufacturing production, and (iv) an index of retail sales in urban areas, for the United States in each case.
 - b) Samples used in index numbers are usually stratified. Why?
 - c) Compare the advantages of random, systematic, and judgment sampling in selecting items for a price index representing a comprehensive list of women's apparel items.
9. If you were to choose a new base period to replace the old 1957-1959 base for federal government indexes, what period of years would you choose? Appraise the merits and drawbacks of this period, according to the four criteria given in this chapter for choice of a base period.
10. a) Convert the American Appraisal Co. index of construction costs, below, to the 1957-1959 average as base.
 - b) Compare the changes in construction costs since 1957, as shown by the Department of Commerce and American Appraisal Co. indexes.
 - c) If in early 1968 the only available construction cost index for 1967 were the E. H. Boeckh figure of 126.0, compared with 122.1 for 1966, use these figures to estimate the American Appraisal Co. index (1957-1959 = 100) for 1967.

CONSTRUCTION COST INDEXES

	U.S. Department of Commerce (1957-1959 = 100)	American Appraisal Company (1913 = 100)
1957	99	663
1958	100	682
1959	102	704
1960	103	722
1961	104	741
1962	107	756
1963	109	780
1964	112	802
1965	116	824
1966	121	867

SOURCE: U.S. Department of Commerce, *Business Statistics*, 1965, p. 52, and *Survey of Current Business* (February 1967), p. 8-9.

11. Find an article in *Monthly Labor Review* or elsewhere reporting on the Bureau of Labor Statistics' five-year program of revising the Consumer Price Index during fiscal 1960–1964. Describe the principal steps in this program and explain how the resulting improvements justify the considerable expense involved.
12. The Ford Motor Company's agreement of September 1958 with the UAW–CIO unions called for a quarterly "cost of living allowance" of approximately 1 cent per hour in straight-time hourly earnings for each 0.5 point change in the Bureau of Labor Statistics Consumer Price Index (1947–1949 = 100) above, but not below, the base index level of 119.1 beginning with 1 cent for index 119.2 to 119.6. (The November 1958 index was 123.7.)

In another case, the H Company reached an agreement with the Metal Workers' Union stating that if the Consumer Price Index increased or decreased by 5 percent or more in any semiannual period, wages would be adjusted upward or downward by the same percent.

Compare the merits of these two agreements as to:

- a) Adjusting wages at all levels by 1 cent per hour for each 0.5 point change in the Consumer Price Index versus adjusting wages by the same *percent* amount as the change in the Consumer Price Index.
 - b) Adjusting wages in little jumps (i.e., quarterly, for each 0.5 point change in the Consumer Price Index) versus big jumps (i.e., semiannually, by 5 percent or more, provided the Consumer Price Index has changed that much).
 - c) Setting a minimum level of wages 4.6 cents an hour below the September 1958 rate, as indicated in the first paragraph, versus adjusting wages upward or downward without limit, in line with the Consumer Price Index.
13. Why is the Bureau of Labor Statistics Wholesale Price Index, excluding farm products and foods, frequently used in place of the All Commodities Index as a measure of general price changes?
 14. If you were the economist of a national chain of drugstores and wished to compare the prices you pay with the Bureau of Labor Statistics Wholesale Price Index:
 - a) Which subgroups of this index would you combine to meet your needs?
 - b) What method, arithmetically, would you employ to combine them?
 15. Is the following procedure appropriate? If not, suggest improvements. In order to allow for changes in the cost of living, a wage contract is set up by the Ajax Machine Tool Company of Houston, Texas, providing that machine tool workers' wages will be adjusted upward or downward each month by 1 cent per hour for each one-point change in the Wholesale Price Index.
 16. What subindex or group of subindexes of the Federal Reserve Monthly Index of Industrial Production is appropriate for comparisons with the physical volume of production of:
-

- a)* A large integrated oil company?
 - b)* A manufacturer of home laundry and kitchen appliances?
 - c)* A household furniture factory?
 - 17. Present a critical analysis of a composite business or economic index of interest to you (other than the Bureau of Labor Statistics price indexes or the Federal Reserve Board Index of Industrial Production), describing its (*a*) purpose, (*b*) method of construction, and (*c*) limitations. (See Selected Readings, below, for sources.)
 - 18. Considering the economic characteristics of your own state or area:
 - a)* List four business indicators that are most significant for this state, giving exact sources.
 - b)* Describe and appraise a general business index published for this state or area.
 - 19. What published indexes or indicators are most appropriate for use in the following situations?
 - a)* You wish to set a price at which to sell your frame house, which you bought new for \$15,000 four years ago.
 - b)* The manager of a wool textile mill is anxious to learn if the expansion in his volume of production over the past 18 months has kept pace with that of the industry.
 - c)* The controller of a gas and electric company needs an adjustment factor with which to revise the basic level of pension payments, set up ten years ago, for the company's retired workers.
 - d)* An agricultural implement manufacturer needs information on recent trends in farmers' operating margins.
 - e)* The president of a chain of department stores desires a monthly measure of changes in consumer purchasing power. He intends to compare this with the sales of his stores.
 - 20. Justify or criticize the following actions. If an action is incorrect, indicate what should be done instead.
 - a)* An oil company economist is asked to compare the growth of his industry since 1935 with that of industry in general. He prepares a ratio chart showing total dollar sales of the oil industry each year, expressed as index numbers on a 1957 base, together with the Federal Reserve Bureau Index of Industrial Production.
 - b)* An executive in Kansas City is offered a job in Cleveland and wishes to compare the cost of living in the two cities. The latest Consumer Price Index is 115.3 for Kansas City and 108.1 for Cleveland. Therefore, he concludes that living costs are somewhat lower in Cleveland.
 - c)* The purchasing agent for a chain of automobile accessory stores who buys his major items direct from manufacturers needs a summary measure of general price changes each month with which to compare his costs. He chooses the Bureau of Labor Statistics Wholesale Price Index for this purpose.
-

- d) A newspaper writer observes that gross national product has increased from \$100 billion in 1940 to \$700 billion in 1966. Therefore, he reports that the nation's output of goods and services has increased sevenfold over this period.

SELECTED READINGS

DOODY, FRANCIS S. *Introduction to the Use of Economic Indicators*. New York: Random House, 1965.

A guide to economic measurement and forecasting, with exercises in the use of major indicators.

FEDERAL RESERVE BOARD. *Industrial Production Measurement in the United States: Concepts, Uses, and Compilation Practices*. Washington, D.C.: Board of Governors of the Federal Reserve System, 1964.

An authoritative treatment of principles and methods of constructing a quantity index.

JOINT ECONOMIC COMMITTEE, U.S. CONGRESS, 1964 *Supplement to Economic Indicators*. Washington, D. C.: U.S. Government Printing Office, 1964.

Contains brief descriptions of the series regularly included in *Economic Indicators* and describes uses and limitations of each.

MOORE, GEOFFREY H., ed. *Business Cycle Indicators*. 2 vols., National Bureau of Economic Research. Princeton: Princeton University Press, 1961.

Twenty articles, and basic data, assessing the principal indicators of short-term business fluctuations in the United States and Canada.

SNYDER, RICHARD M. *Measuring Business Changes*. New York: John Wiley, 1955.

A comprehensive analysis and description of American business indicators.

U.S. BUREAU OF THE BUDGET. *Statistical Services of the United States Government*. Rev. ed. Washington, D.C.: U.S. Government Printing Office, 1963.

Part II describes the principal economic series collected by federal agencies.

U.S. DEPARTMENT OF COMMERCE. *Business Statistics*, biennial supplement to the *Survey of Current Business*. Washington, D.C.: U.S. Government Printing Office, 1965 *et seq.*

The "Explanatory Notes to the Statistical Series," referred to in the footnotes of the tables, cover 2,500 monthly or quarterly series.

U.S. DEPARTMENT OF LABOR. *Major BLS Programs—A Summary of Their Characteristics*. Washington, D.C.: U.S. Government Printing Office, 1966.

Contains description of data collection and methods of preparing all of the major Bureau of Labor Statistics series.

19. TIME SERIES ANALYSIS: SECULAR TREND

MODERN BUSINESS and economic affairs are intensely dynamic in nature. "The old order changeth," sometimes with bewildering rapidity, and the analyst must be alert to interpret the significance of the passing scene. The changes are of many types. The long-term growth of industrial production, the residential building cycle, seasonal swings in department store sales, the daily movements of stock prices, and countless other elements in the dynamics of enterprise must be measured and appraised as an aid in understanding the experience of the past and in formulating future policy. The importance of dynamic fluctuations, as opposed to static analysis, is reflected by the fact that the great bulk of data in business and economic publications (e.g., *Survey of Current Business*, *Economic Indicators*) is in the form of time series rather than being primarily classified by size, space, or other qualitative criteria at a given point of time.

TYPES OF BUSINESS FLUCTUATIONS

It is not sufficient for a businessman to observe merely the overall behavior of an economic indicator. There are various factors at work, the combined effect of which produced this result. Suppose a company's sales increased 6 percent over last month. Was this increase attributable to normal growth, a cyclical business boom, a pickup in seasonal demand, or an advertising campaign? What action should be taken as a result? Analysis of the data involves segregation of these factors so that their separate importance can be understood. The first necessity, then, is to know what factors are present in a time series. Next, how can the

effect of each force be measured? Finally, how can it be predicted as an aid to forward planning?

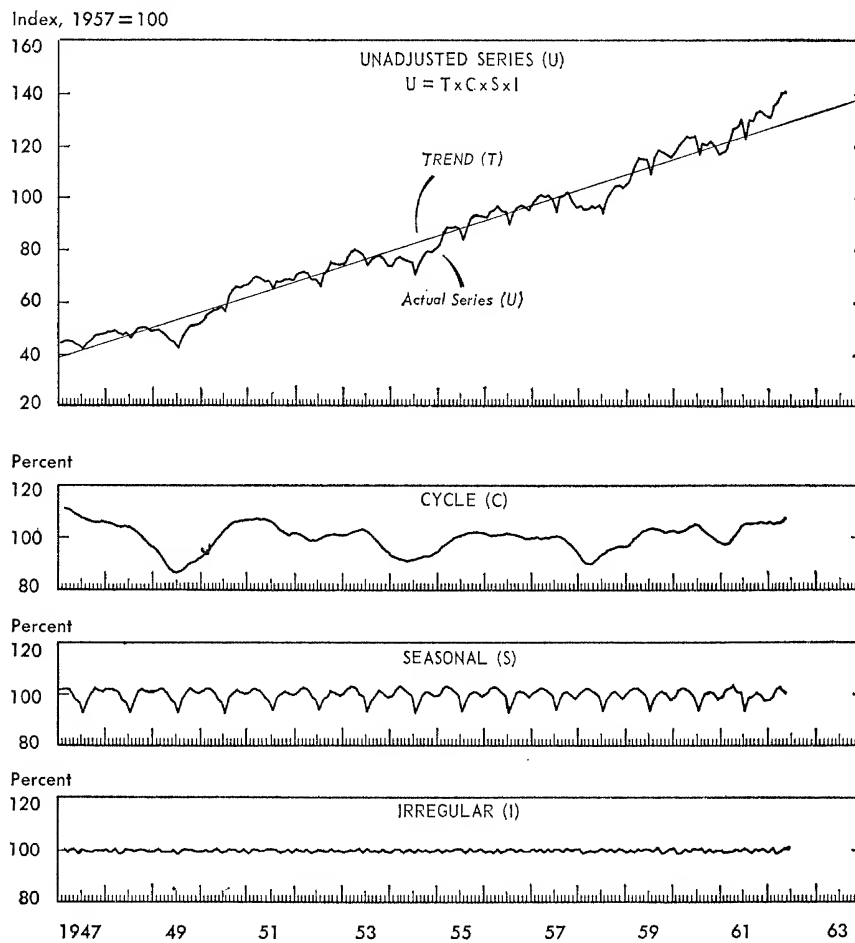
The principal component fluctuations in a time series are as follows:

1. Secular trend.
2. Cyclical fluctuations.
3. Seasonal variation.
4. Irregular movements.

To illustrate, Chart 19-1 shows the monthly production of chemicals over a 15-year period, broken down into a rising trend, the wavelike

Chart 19-1

THE ANATOMY OF A TIME SERIES
PRODUCTION OF CHEMICALS AND RELATED PRODUCTS



SOURCE: Federal Reserve Board index analyzed in *Survey of Current Business* (September 1962), p. 25.

cycles having a period of three to five years, the seasonal movement repeating its pattern each twelve months, and a small irregular residual. The trend value is measured in the original unit of the series (an index number in this case), while the other three components are expressed in percentages. The product of the four components makes up the actual series.

Some time series contain all of the foregoing elements; others contain only some of them. Certain series are so largely controlled by one type of fluctuation that it is easily recognized from the original data. Thus, the production of synthetic fibers and frozen foods have a strong upward trend, durable goods suffer wide cyclical swings, department store sales are predominantly seasonal, and manufacturers' purchased material inventories move irregularly. Usually, however, the several components are not separately recognizable in the original data, but the businessman or economist needs to know the influence of each in order to understand the forces at work and the probable future behavior of the series. Therefore, the analyst's problem in dealing with time series is to identify the components and measure them separately.

The work of analysis can be divided into three parts: (1) fitting a secular trend curve, (2) measuring seasonal variation, and (3) analyzing cyclical-irregular residuals.

This chapter and the next two contain an explanation of the most useful methods for carrying out these three steps in the analysis of time series. In a particular application, only one or perhaps two of the steps may be needed, depending on the importance of the component and the purpose of the study.

SECULAR TREND

Secular trend is the gradual growth or decline of a series over a long period of time. The growth is ordinarily one of physical volume, like biological change; it does not strictly apply to long-term movements in prices, which do not grow in the biological sense. Hence, secular trend analysis usually applies to physical volume series and "deflated" dollar value series, expressed in constant dollars, rather than to dollar value or price series. However, trend curves are sometimes used to describe long-term movements in prices, even though the rational basis of growth is absent.

The tremendous expansion of population and technology in recent decades has stimulated widespread interest in the problem of measuring and predicting economic growth. Long-range planning has become a "must" for progressive companies, and trends must be projected as the

first step in making a complete forecast and in setting a viable goal for future operations. It is particularly important to gauge the growth trends for individual industries and products, since they vary so widely, from the explosive growth of computers to the dismal decline of the railway passenger business. Most industries will also vary in their own rate of growth over a long period of years.

The variations in the nature of the secular trend component can be seen in the three curves of Chart 19-2. Gross national product in constant dollars represents the physical volume of total production; aluminum production typifies a young industry, and bituminous coal, an older one. The data have been plotted on identical ratio scales, and smooth trend curves have been fitted by the National Industrial Conference Board to indicate average growth tendencies. The slopes of these curves show how the percent rates of change differ in each case.

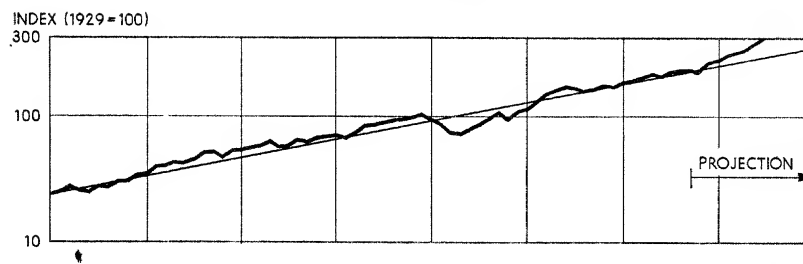
Gross national product has maintained nearly a straight line or uniform percent rate of growth since 1890. Aluminum production, on the other hand, has shot up much more rapidly throughout its short life, although the trend curvature indicates that the rate of growth is slackening. The older bituminous coal industry developed at a more gradual rate from 1890 until World War I; since then it has matured and leveled off. Its course, however, has been steadier than that of aluminum. The three production series therefore exhibit marked differences in (1) shape of trend curve; (2) steepness of curve, or rate of growth; and (3) instability, measured in deviations from the curve. Trend analysis is most useful and reliable when growth is steady and steep and when the deviations about the trend curve are small. In this case the trend curve may even be projected into the future as a forecast if the factors affecting past growth are expected to continue.

The trend types in Chart 19-2 illustrate the industrial application of a useful growth hypothesis popularly called the "law of growth." According to this principle, "If the population is expanding freely over unoccupied country, the percent rate of increase is constant. If it is growing in a limited area, the percentage rate of increase must tend to get less and less as population grows . . ."¹ until it finally levels off as an upper limit is approached. The constant rate of growth is characteristic not only of young industries (e.g., aluminum) but of total produc-

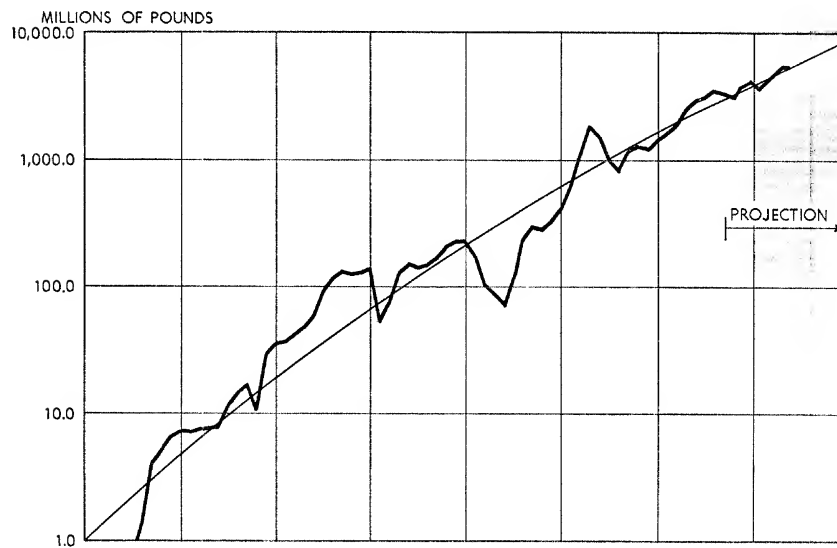
¹ P. F. Verhulst, "Recherches mathematiques sur la loi d'accroissement de la population," *Nouveaux memoires de l'Academie Royale de Sciences et Belles-Lettres de Bruxelles*, Tome XVIII (1845). See also Raymond Pearl and Lowell J. Reed, "On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation," *Proceedings of the National Academy of Sciences* (June 15, 1920), pp. 275-88.

Chart 19-2

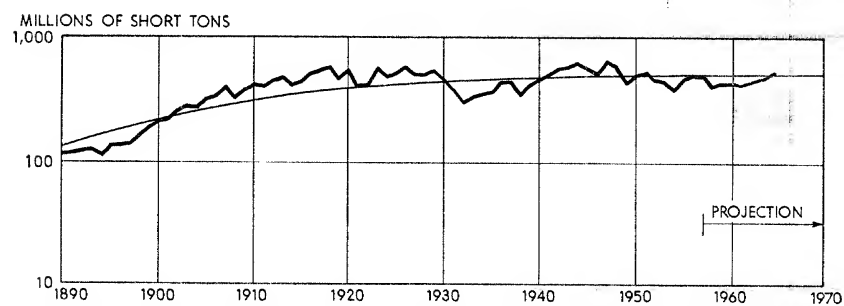
GROWTH PATTERNS IN AMERICAN INDUSTRY, 1890-1965

GROSS NATIONAL PRODUCT
(Constant Dollars)

PRIMARY PRODUCTION OF ALUMINUM



BITUMINOUS COAL PRODUCTION

Source: National Industrial Conference Board, *Growth Patterns: A Reexamination*, pp. 53, 40, 42.

tion (e.g., GNP), which is a cumulation of individual growth curves. The "law of growth" principle will be applied to the measurement of industrial trends later in the chapter.

These examples are sufficient evidence that the growth factor may be described by a simple curve, although it differs for each series. The problem of trend measurement, however, is not merely the mechanical one of fitting a curve to the data; it also requires a knowledge of the background of the industry under consideration. With this knowledge, one can apply methods of time series analysis that are not only mechanically correct but logical as well.

Purposes of Measuring Trend

There are three principal purposes of measuring secular trend:

1. The first purpose is to study the past growth or decline of a series. The secular trend curve describes the basic growth tendency of a product or industry, ignoring short-term fluctuations due to business cycles, seasons, wars, or other causes. The trend curve answers such questions as: Has the company maintained its historic rate of expansion in recent years or is this rate tapering off? Has the company kept pace with its competitors or with the industry as a whole? Is this a "growth" or a stable industry or perhaps a declining one?

2. The second and most important purpose of measuring secular trend is to project the curve into the future as a long-term forecast. If the past growth has been steady and if the conditions that determine this growth may reasonably be expected to persist in the future, a trend curve may be projected over five to ten years into the future as a preliminary forecast. Then regression analysis can be applied (Chapters 22–24), and a qualitative study of other factors, such as business cycles and specific demand and supply conditions, should be made to modify the trend forecast.

A long-term forecast is desirable in making a decision to take a job with a given company or to invest in its stock. It is even more essential in the management's decision to expand its plant, develop a new product, or enter a new regional market in order to justify the capital expansion. The projection of trend curves into the future is subject to considerable error and is deplored by many because of its inexactness and dependence on subjective judgment. Nevertheless it is a necessary expedient, since any major business decision affecting future operations involves a forecast, whether explicit or implicit. In the effort to avoid explicit forecasting the assumption is too often made that present levels will continue unchanged. In a dynamic economy such as ours, this

assumption is apt to lead to poorer planning than with the use of very crude extensions of past trend curves as projections.

3. The third purpose of measuring secular trend is to eliminate it, in order to clarify the cycles and other short-term movements in the data. A steep trend may obscure minor cycles. Dividing the data by the trend values yields ratios which make the curve fluctuate around a horizontal line, thus bringing the cycles into clear relief.

However, these cyclical relatives may be affected arbitrarily by the type of trend curve used and the period to which it is fitted. Also, cycles can usually be discerned without trend adjustment, since the trend component has rarely dominated short-term cyclical-irregular movements in recent times. Hence, the trend is not so often eliminated in current practice as it was formerly. Most government indexes of business activity, for example, are expressed as percentages of some base, such as $1957-59 = 100$, rather than as percentages of the trend values. These indexes show secular growth as well as short-term fluctuations.

The particular purpose of measuring trend affects the choice of a trend curve to some extent. (1) In measuring the past growth of an industry, any type of empirical trend curve that best describes the basic pattern of change may be used, although the logarithmic straight line is best for comparing the average percent rate of change in different series. (2) In projecting trend curves, however, the trend must provide a rational preview of future tendencies as well as fitting the past data. Hence, a decreasing rate of growth curve is often preferable. (3) Finally, in measuring the trend in order to eliminate it, any type of empirical curve that approximately bisects the cycles may be used.

Period of Years Selected

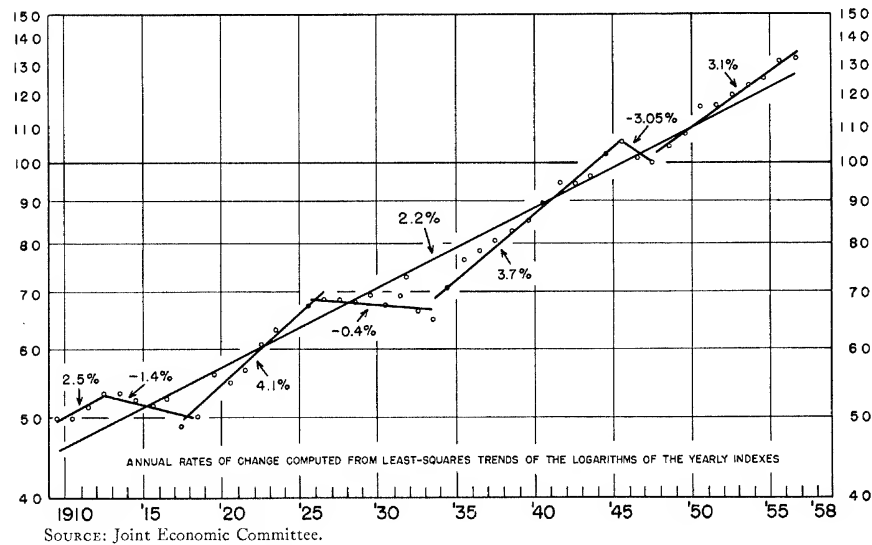
The following rules should be observed in selecting the period of years to be used in fitting a trend curve:

1. The period should be as long as possible, preferably at least 15 years. In a long period the trend curve is but little affected by short-term episodes such as wars and depressions, whereas in a short period a trend measurement may be distorted by these factors.
2. If the nature of a product or industry is abruptly changed by war, the introduction of a new product, or some other fundamental force, the series should be broken at this point and separate curves fitted to each segment. An examination of the graph of the data will be helpful in revealing such changes.

3. Each end of the series should represent the same phase of the business cycle. Thus, if recent years are prosperous, the series might go back to the postwar year 1947 to begin with a prosperous period. If the series began in 1932, the trend line would be tilted upward by the depression at the beginning and prosperity at the end of the period so that it would exaggerate the true basic growth.

Chart 19-3

ANNUAL RATES OF CHANGE IN OUTPUT PER MAN-HOUR
IN THE TOTAL PRIVATE ECONOMY
(1947 = 100)



Serious errors have occurred through fitting trend curves to short periods of years dominated by cycles and other temporary disturbances. In the late 1920's, "trend" curves were fitted from the first postwar year, 1919, through the following decade, a period dominated by the expansion phase of a major cycle. These trends were then projected forward to produce the illusory errors of the "new era." Conversely, pessimistic errors were made in the next decade by fitting curves over periods extending from the prosperous 1920's to the depressed 1930's, thus creating the illusion of a mature or stagnant economy.

Chart 19-3 shows trends fitted to various periods of years in output per man-hour, an important factor determining "productivity" or "improvement factor" increases in wage-rate contracts. Over short periods the average "trend" has varied from a growth of 4.1 percent per year to

a decline of more than 3 percent. In particular, the United Auto Workers have cited the average annual growth of over 3 percent since 1947 to support their demands for future wage-rate increases. On the other hand, the long-term growth since 1909 has averaged only 2.2 percent per year, according to the Joint Economic Committee statisticians.

Price Deflation

Many series on the volume of sales, production, and other economic activities are available only in the form of dollar values. These values are affected not only by the physical quantity of goods involved but also by their prices, and prices have varied widely over the years. For many purposes it is necessary to know how much of the dollar value changes represents a real change in physical quantity and how much is due to mere markups or markdowns in price tags. Physical quantities may be estimated by dividing the dollar values by the prices of the goods represented to eliminate the effect of price changes. (Price data are widely available.) That is, since value equals price times quantity, then value divided by price equals quantity. This adjustment is called price deflation or expressing a series in terms of constant dollars.

For example, suppose the sales in a shoe department increase from \$10,000 in April to \$10,450 in May. What was the change in physical volume? If we ascertain that the average price of shoes increased from \$10 to \$11 a pair in this period, we may divide the value by the price to learn that there has been an actual decline in shoes sold from 1,000 to 950 pairs, as shown below:

DEFLATION OF SHOE SALES

	<i>April</i>	<i>May</i>
1. Dollar sales	\$10,000	\$10,450
2. Average price per pair	\$ 10	\$ 11
3. Estimated number of pairs sold ($1 \div 2$)	1,000	950

Similarly, money wages may be deflated to find "real" wages, that is, wages in terms of the actual goods and services which can be purchased for a given amount of money.

The deflating process is a very simple one; the major problem is the selection of the proper price index. The rule to be followed is "Use an index number computed from the prices of the commodities whose values are to be deflated." For example, hardware store sales should be deflated by an index of hardware prices, not by a general price index.

In deflating dollar values that represent a variety of commodities, an appropriate price index may be pieced together from available sources

to represent this particular "mix." For example, an investor may desire to study the long-term growth of Sears, Roebuck & Company. The secular trend curve should be fitted to the physical volume of sales, since the price changes reflected in dollar sales follow no consistent pattern and merely obscure the real growth. The dollar sales therefore must be divided by a price index of the goods sold by the company.

Such an index might be constructed by pricing a sample of important items sold by the store and weighting these prices by the sales volume of the departments represented. It is simpler, however, and adequate for the purpose, to use existing retail price indexes. The Consumer Price Index itself is not suitable, since it contains elements such as foods, rents, and personal services not sold by the store; but the apparel and house furnishings components of this index may be appropriate. An analysis of Sears, Roebuck sales indicates that roughly half the sales are in apparel and other soft goods, one third in house furnishings and appliances, and one sixth in farm implements and other hard goods. We may therefore weight the Consumer Price Index apparel component one half, the house furnishings component one third, and the Department of Agriculture index of farm machinery prices (excluding motor vehicles) one sixth to get a combined price index appropriate for Sears, Roebuck sales.

To construct this index, the farm price index was first converted from its 1910–1914 base to a 1957–1959 base, as described in Chapter 18, for comparability with the other two indexes. Then, each of the three indexes was multiplied by its weight, and the results were totaled for each year to get the composite price index on a 1957–1959 base. Finally, it was thought desirable to express sales in terms of 1965 prices—since these are more up to date than the price levels of 1957–1959—so the whole series was divided by the 1965 index of 105.0 percent to yield the price indexes on a 1965 base shown in Table 19–1. Dividing reported net sales by this index gives deflated sales (actually "inflated" in this case).

Chart 19–4 compares the actual and deflated sales along with the price index on a ratio grid. The physical volume of business has increased more gradually than reported sales because of price inflation. Furthermore, much of the apparent gain in sales from 1947 to 1951 was due to price markups, whereas nearly the entire rise in sales from 1961 to 1965 represented a real increase in physical volume, since prices were fairly stable during this period. Note that the use of the 1965 base in the price index brings the two sales curves together in this year. If a 1957–1959 base had been used, the deflated sales curve would

Table 19-1

SEARS, ROEBUCK ANNUAL NET SALES, 1947-1965

Year*	Net Sales [†] (Millions of Dollars)	Price Index [‡] (1965 = 100)	Deflated Net Sales [§] (Millions of 1965 Dollars)
1947.....	1,982	82.6	2,400
1948.....	2,296	88.7	2,589
1949.....	2,169	86.9	2,496
1950.....	2,556	86.8	2,945
1951.....	2,657	95.1	2,794
1952.....	2,932	94.1	3,116
1953.....	2,982	93.6	3,186
1954.....	2,965	92.8	3,195
1955.....	3,307	91.9	3,598
1956.....	3,556	92.9	3,828
1957.....	3,601	94.7	3,803
1958.....	3,721	95.0	3,917
1959.....	4,036	96.1	4,200
1960.....	4,134	97.2	4,253
1961.....	4,268	97.7	4,368
1962.....	4,603	97.9	4,702
1963.....	5,116	98.5	5,194
1964.....	5,740	99.2	5,786
1965.....	6,390	100.0	6,390

* Fiscal years beginning February 1.

† Total net sales less discounts, returns, and allowances, including outside sales by subsidiaries. SOURCE: Stockholders' reports.

‡ Constructed from U.S. Department of Commerce Consumer Price Index for apparel (weight $\frac{1}{2}$) and house furnishings (weight $\frac{1}{4}$) plus U.S. Department of Agriculture index of prices paid by farmers for farm machinery (weight $\frac{1}{4}$), adjusted to 1965 base.

§ Net sales divided by price index times 100.

have been lowered to match the other curve in these years, but its slope would not have been altered. Several types of secular trend curves will be fitted to the deflated sales in the next section.

METHODS OF MEASURING TREND

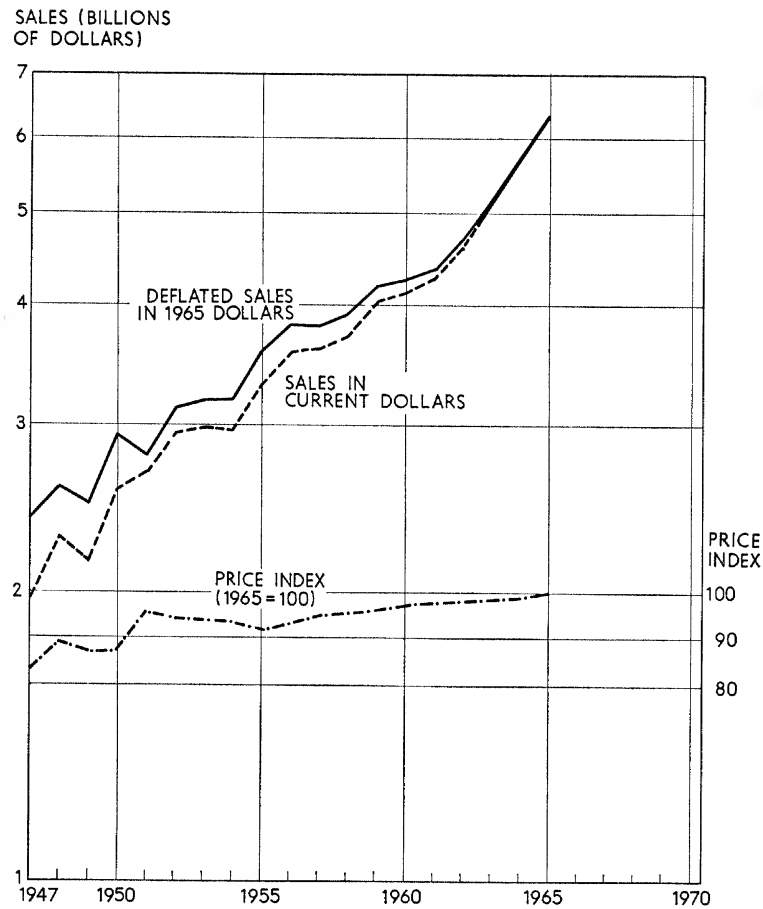
A secular trend curve may be fitted to a series of data by means of (1) a graphic "freehand" fit, (2) the method of selected points, or (3) the method of least squares. These will be described in turn. In each case the statistical technique must be supplemented by a knowledge of the economic forces involved and the rational nature of the growth factor represented.

Annual data are ordinarily used in secular trend analysis, rather than quarterly or monthly figures, because short-term movements are usually insignificant in measuring the broad sweep of an industry's growth or decline and because the use of such detailed data involves much extra

Chart 19-4

SEARS, ROEBUCK ANNUAL NET SALES, 1947-1965

(IN CURRENT AND 1965 DOLLARS)



work. However, the methods applied in this chapter to annual data can be easily adapted to quarterly or monthly figures if desired.

The series should first be plotted on a graph to provide a visual picture of the fluctuations in the data, and later the trend curve and the reasonableness of the fit. The arithmetic scale is somewhat easier to plot and simpler for the reader to understand than the ratio scale. The arithmetic vertical scale is also appropriate for fitting trend equations to the natural values of the data by least squares (to be explained below).

For trend analysis in general, however, it is recommended that the data be plotted on a ratio scale, since this grid shows the two most

important types of trend curves in their simplest form: (1) The exponential curve, with a constant percent rate of growth, appears as a straight line. This logarithmic straight line characterizes many young industries and affords easy comparison of average rates of change in different series. (2) The "growth" curve, with a decreasing rate of gain, appears as a simple curve bending over to the right, as in Chart 19-2, rather than as an elongated S on an arithmetic scale.

Graphic "Freehand" Measurement

The simplest method of fitting a trend curve is to draw it through the center of the plotted data by inspection.² If the general tendency of the data roughly follows a straight line, a transparent ruler or a piece of string may be used to locate the approximate central trend. If the trend is curved, a large-size transparent French curve or an engineer's flexible spline rule may be used. The term "freehand" is applied to any non-mathematical curve in statistical analysis, even when it is constructed with the aid of drafting instruments.

The trend line or curve should be drawn through the graph of the data in such a way that the areas above and below the trend are equal. They should be exactly equal for the series as a whole and approximately equal for the first half and last half of the series separately and as far as possible for each major cycle. That is, the vertical deviations of the data above the trend line must total the same as the vertical deviations below the line. These deviations may be marked off cumulatively on the edge of a strip of paper, one above the other, for comparison. In Chart 19-5, for example,³ the total vertical deviations ($a + b + c$) below the trend line must equal the total of those above ($d + e$).

Use of Group Averages. The average values of groups of data may be plotted as guide points in drawing a smooth trend curve. These averages may be computed for successive three- or five-year periods, or they may be computed for each cycle, marked off from trough to trough and plotted at the center year of the cycle. The trend is then drawn as a

² For a more precise but detailed method of fitting a straight line, see S. I. Askovitz, "A Short-Cut Graphic Method for Fitting the Best Straight Line to a Series of Points According to the Criterion of Least Squares," *Journal of the American Statistical Association* (March 1957), pp. 13-17.

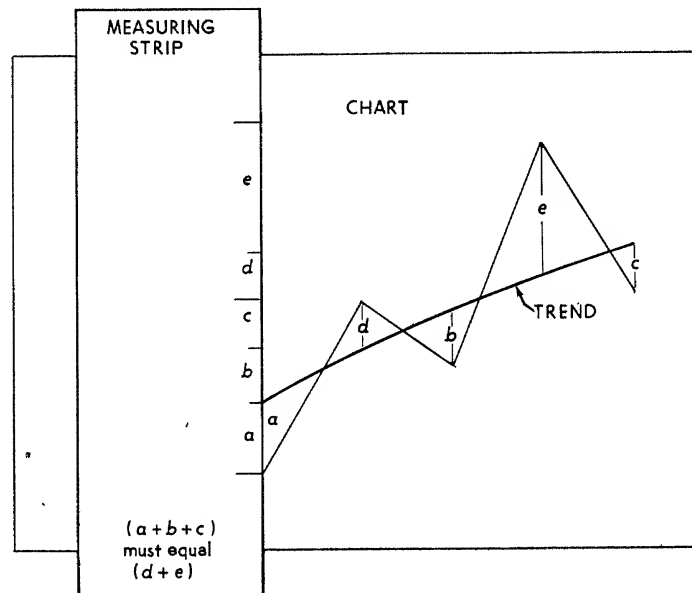
³ This chart also illustrates a graphic method of finding the mean deviation of the data around the trend line, as a measure of cyclical amplitude or instability of growth. Simply cumulate the total deviation ($a + b + c + d + e$) on a paper strip, measure this distance in centimeters or inches, divide by the number of items (5), and lay off the average distance on the vertical scale of the chart to find the mean deviation. On a ratio scale, lay off the average distance above a base line as 100 percent, as described on page 57. If it comes to 108.5, the mean deviation is $108.5 - 100 = 8.5$ percent. Do not read off the total deviation on a ratio scale.

smooth curve between the plotted averages, but not necessarily through each one.

An Example: Fitting and Projecting Graphic Curves. Chart 19-6 shows two secular trend curves fitted by the graphic method to Sears, Roebuck deflated sales from 1926 to 1956. Sales for the next nine years, 1957-65, have then been plotted as a check on the validity of the trend projections that might have been made in 1957 as long-range fore-

Chart 19-5

CHECKING THE FIT OF A FREEHAND TREND CURVE



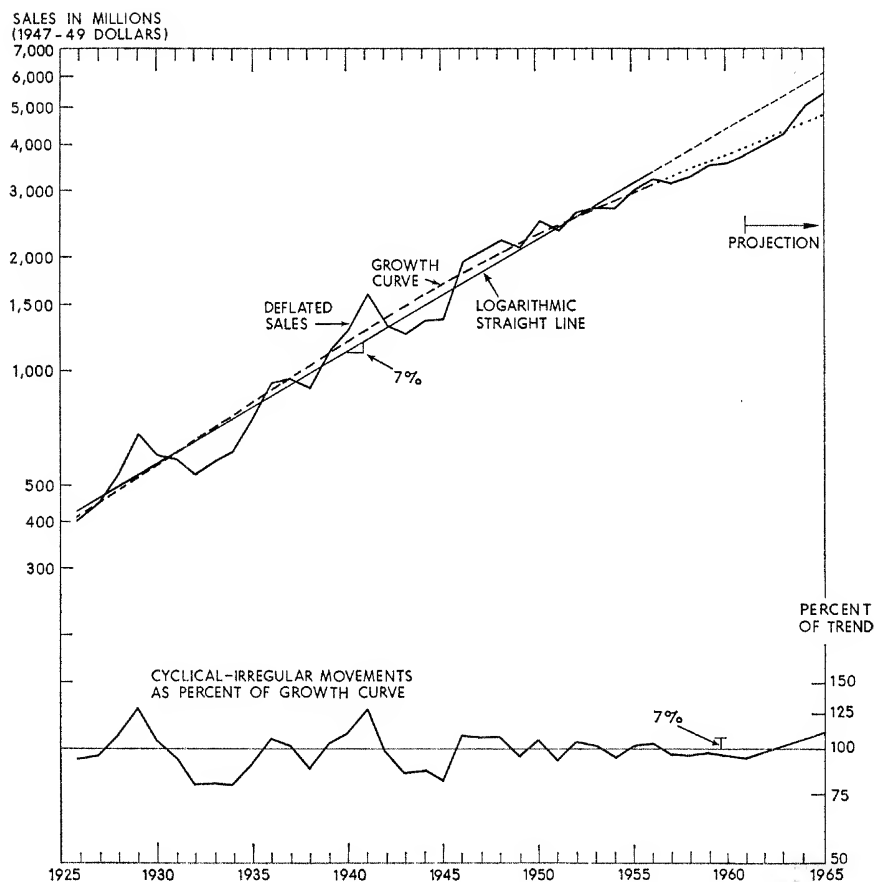
casts. The ratio scale is chosen because the percent rate of growth has been nearly constant during this period, and so it can be represented by a simple straight line, whereas the trend would curve up more and more steeply on an arithmetic chart.

The period of years is long enough so that the trend growth dominates the short-term cyclical-irregular movements. This period also balances the high-level prosperity levels of 1926-1929 and 1952-1956 at its two extremes. Finally, it represents the entire era of the company's expansion in urban department stores, the first one of which was established in 1925.

Since the general growth tendency is nearly linear, a "logarithmic

Chart 19-6

FREEHAND TRENDS FITTED TO SEARS, ROEBUCK
DEFLATED SALES, 1926-1956, AND PROJECTED TO 1965



straight line" has been drawn through the data with a transparent ruler so as to bisect approximately each of the major cycles, as far as possible. Then the vertical deviations above and below the line have been cumulated and the line adjusted slightly to equalize the sum of these deviations for the two halves of the series.

The average annual rate of growth has then been measured as follows: The vertical rise in the trend line in any year (see 1940-1941 in Chart 19-6) has been laid off by dividers on the right-hand percent scale of the chart. This distance extends from 100 percent upward to 107 percent, indicating an average growth of 7 percent per year in

deflated sales over this period. This rate may be compared directly with that in deflated sales of other stores or real personal income, if desired.

The graphic measurement of average growth rate is subject to errors in drawing the slope of the trend line and in reading the result off the chart. The error in slope is small, however, if the trend is linear and the deviations from the trend line small. The error in reading values from a chart is also small if the curve is drawn to a large vertical scale.

The straight line indicates that Sears, Roebuck has expanded at a fairly sustained rate over this 30-year period, although some flattening out is evident after 1947. A "growth" function therefore has been drawn with a French curve to embody a decreasing rate of gain. This curve is higher in the middle and lower at the ends than the straight line. In this case, the growth curve appears to describe the trend of sales somewhat better than the straight line, particularly after 1953. The growth curve may also be preferable for long-term projection into the future, since it follows the retardation-of-growth principle characteristic of many industries.

A logarithmic straight line may be projected for a limited period—say five or ten years—since the rate of expansion may be nearly constant for such a period, and the troublesome problem of curvature is avoided. In the very long run, however, the logarithmic straight line becomes too optimistic since it increases indefinitely at a geometric rate.

The 1957–1965 sales plotted on Chart 19–6 show how the trend projections would have worked out for these years. The extended growth curve predicted the average rate of increase in sales fairly well, while the straight line was consistently too high, as it had indicated it might be, by rising above the actual curve in 1954–1956. On the other hand, a logarithmic straight line fitted only to the postwar years 1947–1956 would have forecast 1957–1965 sales reasonably well. This trend type is fitted by least squares to the postwar years later in the chapter. Of course, trend projections do not forecast cyclical and irregular fluctuations, such as the 1962–1965 boom and the company's expansion in new stores. These factors must be analyzed separately.

Eliminating Trend. The growth component of Sears, Roebuck sales may be eliminated graphically on the ratio chart for the purpose of isolating cyclical-irregular movements as follows: Draw a horizontal line at some convenient level away from the original curve—say opposite the lower printed number 2. Then mark a percent scale with 50, 100, and 150 percent opposite the printed scale numbers 1, 2, and 3, respectively. Caption this scale "Percent of Trend." Now take the *vertical* distances from each point to the original trend (the growth curve in

Chart 19-6) with a divider or paper strip, and lay these distances off in the same years above and below the horizontal 100 percent line. Connect these points with straight lines.

The resulting curve represents the cyclical-irregular movements in sales, since the trend is eliminated or flattened out. (There are no seasonal fluctuations in annual data.) The sales are now "adjusted for trend" or expressed as percentages of the trend values. This graphic adjustment is a short-cut method of dividing the sales data by the corresponding trend values and plotting the results.

The cyclical peak in 1929, the depression trough in 1932-1934, the 1941 peak, the period of war shortages, and the mild postwar cycles are all clearly shown. The cyclical levels at the ends of the series, however, are somewhat uncertain, since the trend curve has a larger error where nearby past or future data are not known.

Graphic versus Mathematical Methods. Graphic "freehand" methods in statistical analysis have three major advantages over mathematical computations:

1. They usually save time and labor. For this reason they are widely used in business analysis where approximate results must be obtained in the minimum time.
2. Graphic curves are more flexible than rigid mathematical functions and, hence, may fit the data more closely. In Chart 19-7, for example, a Gompertz curve has been fitted mathematically to the output of Portland cement from 1890 to 1950. This is a fairly good fit, but the curve is clearly too high from 1893 to 1900, too low from 1905 to 1915, and too nearly horizontal after 1946. A freehand trend drawn by inspection with a French curve (dashed line) appears to be a better fit during these periods (and a better projection from 1951 to 1965).
3. Graphic methods afford a continuing picture of successive steps in analysis. Such a picture aids the observer in planning operations and judging the results. It also provides a visual aid in teaching or explaining the method to others.

Graphic methods, however, also have three major disadvantages:

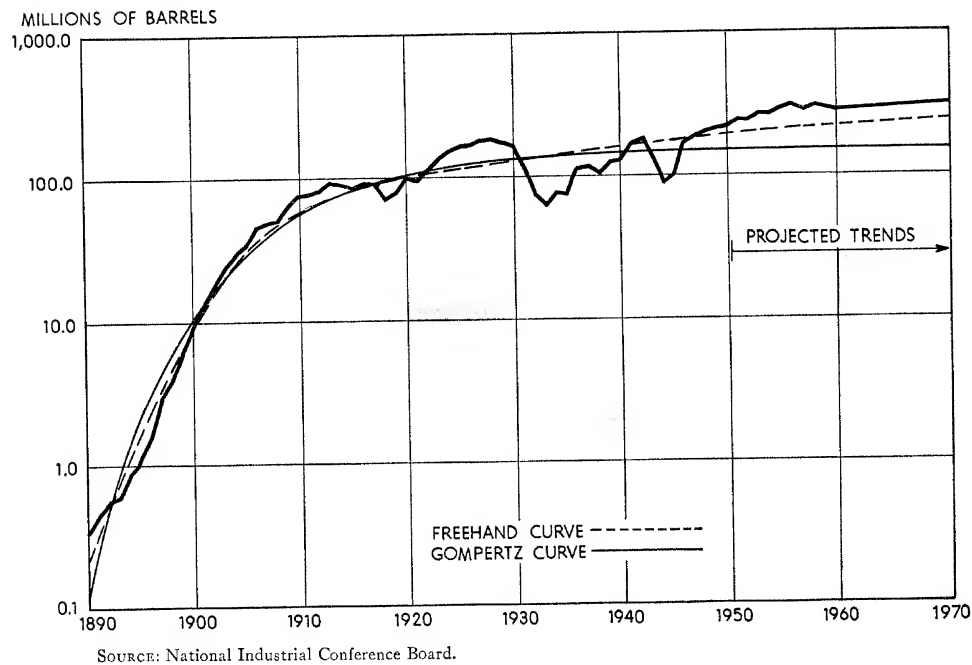
1. They reflect the subjective errors of the analyst. His personal bias, mistakes in judgment, and optical errors all affect the results. However, mathematical techniques, too, require the analyst to

choose the type of equation and period of years to be used. Mathematical methods are no substitute for personal judgment.⁴

2. Because of the subjective element in graphic methods, a skilled analyst is required to draw curves with reasonable accuracy. The amateur may be led astray. Also, where computers are available, repetitive mathematical calculations can be performed rapidly. (Electronic computer programs are available to fit any type of

Chart 19-7

FREEHAND AND GOMPERTZ CURVES FITTED TO OUTPUT
OF PORTLAND CEMENT



SOURCE: National Industrial Conference Board.

polynomial trend by least squares, as a special case of regression analysis. The latter is described in Chapter 24.)

3. Mathematical curves can be expressed by formulas that provide the "best" fit according to some stated criterion. Such results have

⁴ As Simon Kuznets puts it: "We must bear in mind the essential uncertainty of the whole process of separation or we shall be unduly influenced by mechanical methods of fitting. The method of least squares may save the investigator the trouble of decision in fitting to selected points and may seem more objective in the sense that identical results will be reached by different investigators. But mechanical arbitrariness is no whit better for being mechanical, and the method of least squares does not assure satisfaction of the two most obvious criteria of goodness of fit; namely, the balance and the minimizing of relative deviations from trend within each cycle." *Secular Movements in Production and Prices* (New York: Houghton Mifflin, 1930), p. 62.

at least the appearance of greater exactness than do hand-drawn curves and, hence, may carry more conviction with the reader.

Graphic and mathematical methods may be used in combination to utilize the advantages of each. A graphic trend curve, for example, can be drawn to establish its general location and shape; then an appropriate mathematical equation can be selected for more objective measurement. The graphic curve also serves as a rough check on the accuracy and reasonableness of the mathematical equation. In a research department, the director of research can sketch out a preliminary curve graphically, then set up the program for the proper mathematical computations, and finally check the results against his own curves.

The Method of Selected Points: Growth Curves

"Growth" curves may be fitted either graphically, as described above, or mathematically to three selected points. (The equations of these curves are too complex to be easily fitted by the least-squares method described in the next section.) These curves are useful for representing both past trends and probable future tendencies, since they embody the rational "law of growth" principle described above. That is, an industry or population tends to grow at a nearly constant percent rate during its youth; but as it matures, this rate tends to diminish.

There are several types of growth curves—the logistic (Pearl-Reed) and Gompertz being the most common⁵—but all have the general characteristics illustrated in Chart 19-8. Here the same logistic curve is plotted on an arithmetic scale in panel A and a ratio scale in panel B. During the period shown, the curve rises from 1 to 99 and approaches an upper limit of 100.

The elongated S curve in panel A shows the growth of a typical industry or product in absolute units. The first stage is one of experimentation and slow initial growth. Second, there is a period of rapid exploitation of the product, and third, a leveling off of growth with maturity and saturation of demand. The relative age of different industries may be determined by locating them on this curve. Thus, the electronics and atomic energy industries would be located near its beginning; flour milling and railroads, near the saturation level.

The same curve plotted on a ratio scale (panel B) is simpler in form,

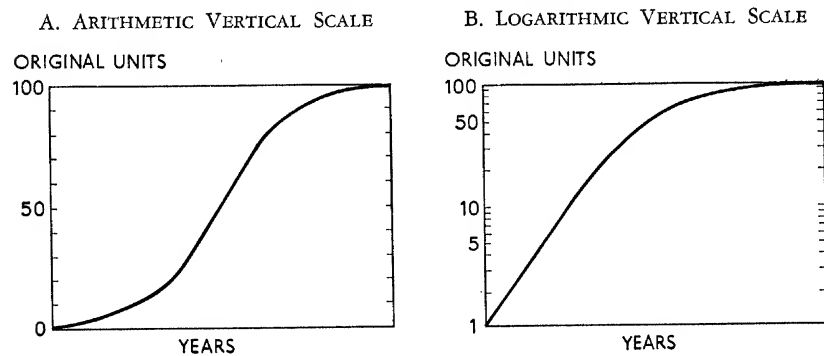
⁵ "The simple logistic and Gompertz curves, mostly the former, describe well the long-term movements of growing industries, and, with certain modifications, those of declining industries." Simon S. Kuznets, *Secular Movements in Production and Prices*, p. 197.

being concave downward throughout its length. This is the grid that best illustrates the growth principle of a nearly constant percent rate of change at first, followed by smaller and smaller percent gains as the industry ages. The data should be plotted on a ratio grid in any of the methods described below.

Before fitting a growth curve, two conditions should be satisfied: (1) The process represented should have the characteristics of biological growth to justify the use of this curve on logical grounds. Prices, ratios, business failures, or unemployment series would not qualify. (2) The data, when plotted on a ratio scale, must show a declining rate of growth or decline (i.e., must tend to flatten out) empirically, like this: growing series, \nearrow ; declining series \searrow . Otherwise, a growth function cannot be fitted.

Chart 19-8

THE LOGISTIC GROWTH CURVE



A growth curve may be fitted to a series of data in any of three ways:

1. The graphic "freehand" method has already been applied to Sears' sales and cement production (Charts 19-6 and 19-7). Plot the data on a ratio scale, and with a French curve draw a smooth trend that bends toward the horizontal, as in panel B, to fit the plotted points. As indicated, this method is easy and flexible, but it involves errors of personal judgment, particularly in projections into the future.
2. In the mathematical method the appropriate type of equation is fitted to three points which are selected at equal intervals of time to represent typical stages of early, middle, and recent development. These points are usually averages of several years to iron

out cyclical-irregular influences. Three constants must then be computed to determine the trend equation. The procedure will not be presented here.⁶

Charts 19-2 and 19-7 show Gompertz curves fitted mathematically by the National Industrial Conference Board to four series for more than a half century through 1958 (1957 for GNP). We have plotted the actual data through 1965 and extended the trend curves to test their validity as projections. GNP exceeded its trend extrapolation in the 1960's, but aluminum was surprisingly close to the trend curve. Coal and cement lagged, though coal reached its trend projection in 1965.

3. A short-cut method may be used to fit a growth curve to three selected points, using a nomograph to determine the upper limit and a special grid on which the growth curve can be drawn as a straight line.⁷ The result approximates that of the corresponding mathematical method.

A growth curve fitted to three points may be a poor fit if the analyst errs in his choice of the appropriate type of equation, the period of years covered, or the three points he believes to be typical. Different equation types and different selections of three points therefore may be tried to achieve an optimum fit in either the mathematical or the short-cut method, though such experimentation is easier in the latter case.

The Method of Least Squares

A simple polynomial equation can often be used to describe the secular trend of a time series. Such an equation provides an objective and concise expression for the growth or decline of the series, but the form of the equation places certain limitations on the possible shapes of the fitted curve.

After the general equation of the trend curve has been selected, the curve is fitted by determining the constants (e.g., a and b in the equations below) so as to obtain the particular curve of the chosen type which fits best. Goodness of fit can be judged in several ways. For

⁶ See F. E. Croxton and D. J. Cowden, *Practical Business Statistics* (3d ed.; Englewood Cliffs, New Jersey: Prentice-Hall, 1960), Chap. 38, for a description of mathematical methods of fitting logistic, Gompertz, and modified exponential curves.

⁷ See William A. Spurr and David R. Arnold, "A Short-Cut Method of Fitting a Logistic Curve," *Journal of the American Statistical Association* (March 1948), pp. 127-134; Eugene A. Rasor, "The Fitting of Logistic Curves by Means of a Nomograph," *ibid.* (December 1949), pp. 548-553; and Jack Sherman and W. J. Morrison, "Simplified Procedures for Fitting a Gompertz Curve and a Modified Exponential Curve," *ibid.* (March 1950), pp. 87-97.

example, one might like to have the average trend values equal the corresponding averages of the data not only for the series as a whole but also for selected parts (e.g., halves or thirds), or one might prefer to have the fitted curve pass through certain key points, such as cycle averages.

The most widely used criterion is that of *least squares*. This criterion states that the best-fitting curve of a given type is the one from which the sum of the *squared* deviations of the data is least. Hence, the "method of least squares." The deviations are measured *vertically* from the trend line, not perpendicularly. This criterion also requires that the sum of the deviations of the data (Y) above the trend line (Y_c) must equal the sum of the minus deviations below the line, so that the total deviations equal zero.

The method of least squares is applied here to the arithmetic straight line, the parabola, and the logarithmic straight line in turn. The sum of the squared deviations from the least-squares straight line is less than that from any other straight line. Similarly, the sum of the squared deviations from the least-squares parabola is less than that from any other curve described by a polynomial in X and X^2 . Since the logarithmic straight line is fitted to the logarithms of the data, the sum of squares of logarithmic deviations is minimized. These usually correspond closely to percent or relative deviations from trend rather than to the absolute deviations.

The method of least squares is most appropriate for data having a uniform variance of deviations along the trend line, few extreme deviations, and deviations that are independent of each other, especially in adjacent periods. These conditions do not hold in time series. The deviations from trend are cyclical-irregular rather than random. Hence, one should attribute no special virtues to the method of least squares for fitting trends except simplicity from a practical point of view.

No matter what method is used to fit a trend, the equation type should be capable of describing the basic tendency of the series. Straight lines are often fitted to series having curved trends, with ridiculous results. Even if a straight line or parabola fits the past growth accurately, it is a purely empirical description and will not necessarily fit future growth. There should be some logical justification for curves used in forecasting, such as the tendency of many industries to grow at a constant percent rate in their youth and at a decreasing rate as they mature. These tendencies are described by logarithmic straight lines and growth curves, respectively.

Arithmetic Straight Line. The general equation of an arithmetic

straight line trend is $Y_c = a + bX$, where Y_c is the computed or trend value of the time series Y in the year numbered X . The constant a is the value of Y_c when $X = 0$, and the constant b is the slope of the trend line—the change in Y_c per unit change in X . In the method of least squares, the trend line is fitted by finding the values of a and b that minimize the sum of the squared deviations from the trend line. To do this, two conditions called the normal equations must be satisfied, since there are two constants in this equation. These equations are

$$\begin{aligned}\Sigma Y &= Na + b\Sigma X \\ \Sigma XY &= a\Sigma X + b\Sigma X^2\end{aligned}$$

where N is the number of items in the series.

The variable X can be measured from any point in time as the origin, such as the first year of the series. It is easier, however, to choose the origin at the *midpoint* in time because the negative values of X in the first half of the series balance out the positive values in the second half, so that $\Sigma X = 0$. In other words, the time variable is measured as a deviation from its mean. Accordingly, X is changed to the small letter x , where $x = X - \bar{X}$. Since $\Sigma x = 0$, the terms containing ΣX drop out of the normal equations, which become

$$\begin{aligned}\Sigma Y &= Na \\ \Sigma xY &= b\Sigma x^2\end{aligned}$$

Solving these equations for a and b ,

$$\begin{aligned}a &= \frac{\Sigma Y}{N} \\ b &= \frac{\Sigma xY}{\Sigma x^2}\end{aligned}$$

where x is measured from the middle year as origin. Here, the constant a is the arithmetic mean of the series and b is a simple ratio.

A straight line trend can now be fitted by the method of least squares as follows:

1. Set up a table with columns for the year (x), the value of the time series (Y), the product xY , and x^2 for each year. (The column for x^2 may be omitted, if desired, by looking up Σx^2 in Appendix K.)

2. Add the columns and substitute the totals ΣY , ΣxY , and Σx^2 in the above formulas to find the constants a and b of the trend equation $Y_e = a + bx$.
3. Take any two values of x (preferably rather far apart), find the value Y_e from the trend equation in each case, plot the corresponding points, and draw a straight line through them. This is the trend line.

If there is an *even* number of years in the series, the x origin must be placed midway between the two middle years in order to make $\Sigma x = 0$. From this origin it is $\frac{1}{2}$ year to the middle of the next year, $1\frac{1}{2}$ years to the middle of the following year, and so on. In order to avoid fractions, therefore, let the x unit equal six months. Then mark the x values of the years following the origin 1, 3, 5, 7 . . . , and the x values going back from the origin -1 , -3 , -5 , -7 The computation proceeds as above, and Σx^2 may be found in Appendix K. Then a is again the trend value at the origin, but b is the increase in the trend in six months rather than in a year.

Another way to simplify calculations for an even number of years is to drop or add a year at the beginning of the period to make the number an odd one. This change will have little effect if the series is long enough for adequate trend measurement.

The trend values (Y_e) can be listed for each year, if desired, by computing the value for the first year and adding the b value successively on a calculating machine to get the other trend values. Note that $\Sigma Y_e = \Sigma Y$ as a check.

Occasionally it is desired to eliminate trend, in order to clarify cyclical-irregular movements. To do this, compute and plot Y/Y_e for each year. As in other statistical adjustments, dividing by a factor ($Y_e = \text{trend}$) eliminates the influence of that factor.

As an example, an arithmetic straight line is fitted to Sears, Roebuck deflated sales in Table 19-2. In our graphic analysis of sales trends from 1925 to 1965 (Chart 19-6), we noted that the rate of growth in Sears, Roebuck sales had declined slightly since 1947. Therefore, we now measure the postwar trend from 1947 to 1965. This 19-year period is long enough for the growth factor to dominate cyclical-irregular influences; also, no abrupt changes have affected Sears, Roebuck sales since World War II (except the short-lived Korean War buying scares); and, finally, the beginning and ending years were both prosperous. Hence, the period of years chosen is a reasonable one.

To compute the trend equation, mark off the x values as integers

Table 19-2

ARITHMETIC STRAIGHT LINE FITTED BY LEAST SQUARES
TO SEARS, ROEBUCK DEFLATED NET SALES, 1947-1965

Year (1)	x (2)	Deflated Sales (Millions) - Y (3)	xY (4)	x^2 (5)
1947.....	-9	2,400	-21,600	81
1948.....	-8	2,589	-20,712	64
1949.....	-7	2,496	-17,472	49
1950.....	-6	2,945	-17,670	36
1951.....	-5	2,794	-13,970	25
1952.....	-4	3,116	-12,464	16
1953.....	-3	3,186	-9,558	9
1954.....	-2	3,195	-6,390	4
1955.....	-1	3,598	-3,598	1
1956.....	0	3,828	0	0
1957.....	1	3,803	3,803	1
1958.....	2	3,917	7,834	4
1959.....	3	4,200	12,600	9
1960.....	4	4,253	17,012	16
1961.....	5	4,368	21,840	25
1962.....	6	4,702	28,212	36
1963.....	7	5,194	36,358	49
1964.....	8	5,786	46,288	64
1965.....	9	6,390	57,510	81
Totals.....	0	72,760	108,023	570

SOURCE: Table 19-1.

from the middle year 1956 as origin, let Y = sales, compute xY and x^2 (or look up Σx^2 in Appendix K), and total these columns. Then

$$a = \frac{\Sigma Y}{N} = \frac{72,760}{19} = 3,829.5 \quad (\text{i.e., the average sales in millions of dollars})$$

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{108,023}{570} = 189.514 \quad (\text{i.e., the average increase per year in millions of dollars})$$

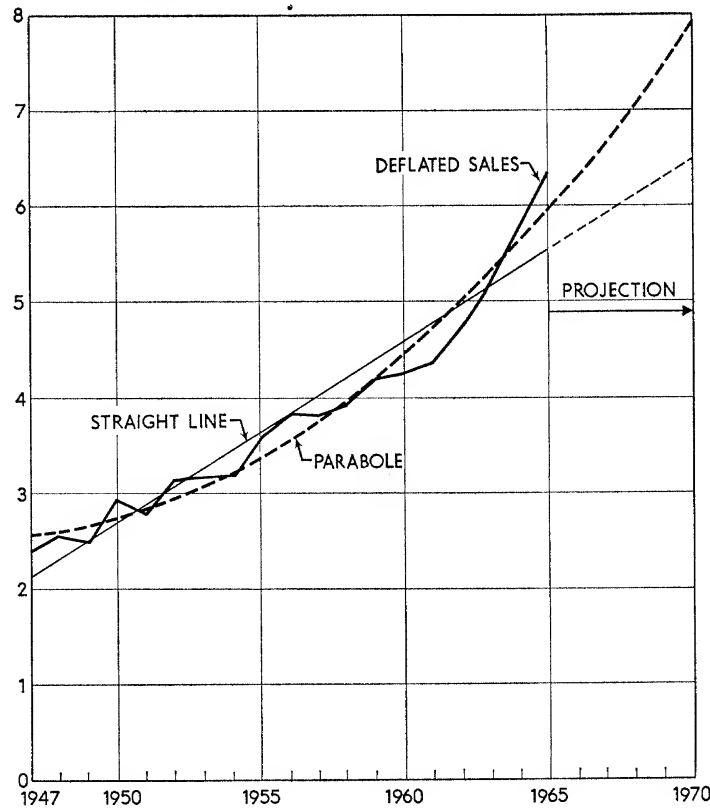
and the trend equation is $Y_c = 3,829.5 + 189.514x$. This equation is plotted in Chart 19-9. It is a poor fit; the line is too high throughout the middle of the series and too low at the ends. Its projection to 1964-1965 falls far below actual sales in those years, and its extension into the past goes below zero in 1935!

The indiscriminate use of the arithmetic straight line is a common error in trend analysis. For example, a large steel company featured this

Chart 19-9

STRAIGHT LINE AND PARABOLA FITTED BY LEAST SQUARES
TO SEARS, ROEBUCK DEFLATED SALES, 1947-1965, PROJECTED TO 1970

SALES IN BILLIONS
OF 1965 DOLLARS



"standard" trend equation in a full-page magazine advertisement to emphasize the growth in per capita production of light steel products since 1901. The result was similar to that in Chart 19-9: The production data curved more and more steeply upward, while the straight trend line touched this curve at only two points and was far below it at the ends. An arithmetic straight line is a valid measure of trend for a series that tends to increase or decrease by constant absolute increments, but it cannot describe the long-term growth of an industry that expands by bigger increments as the industry itself increases in size. A type of trend curve must be chosen that will follow the tendency of a series throughout its course and will pass as nearly as possible through the center of individual cycles.

Parabola. The parabola is more flexible than the straight line as a measure of trend because of its curvature. The general shape of a parabola is that of an automobile headlight reflector, pointing either up or down in its usual form. The values of the data will determine automatically what segment of the parabola will be fitted.

The equation of the parabola which is useful in statistical work is $Y = a + bX + cX^2$, or $Y = a + bx + cx^2$ when the x origin is placed at the middle year. It is called a second-degree equation because X is raised to the second power. This equation contains the three constants, a , b , and c , which may be found as follows by the least-squares method: First compute b by the same formula as in the straight line:

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{108,023}{570} = 189.514$$

Then find a and c by solving the following normal equations simultaneously:

$$\Sigma Y = Na + c\Sigma x^2 \quad (1)$$

$$\Sigma x^2 Y = a\Sigma x^2 + c\Sigma x^4 \quad (2)$$

In addition to the totals shown in Table 19-2, we need $\Sigma x^2 Y$ (column 2 \times column 4, not shown in detail) and Σx^4 (from Appendix K). Here, $\Sigma x^2 Y = 2,299,369$ and $\Sigma x^4 = 30,666$. Substituting in the above equations,

$$72,760 = 19a + 570c \quad (1)$$

$$2,299,369 = 570a + 30,666c \quad (2)$$

Multiplying Equation 1 by 30, to equalize the coefficients of a ,

$$2,182,800 = 570a + 17,100c$$

Subtracting this from Equation 2,

$$\begin{aligned} 116,569 &= 13,566c \\ c &= 8.593 \end{aligned}$$

Substituting this value in Equation 1,

$$\begin{aligned} 72,760 &= 19a + 4,898 \\ a &= 3,571.7 \end{aligned}$$

Hence, the equation of the parabola fitted to Sears, Roebuck sales is

$$Y_c = 3,571.7 + 189.514x + 8.593x^2 \quad (\text{origin } 1956)$$

Finally, compute Y_c at three-year intervals and plot as on Chart 19-9. Here, a is the height of the curve at the origin (but not the arithmetic mean); b is the slope of the curve at this point only; and c determines the amount and direction of curvature. The numerical values are in millions of dollars at 1965 prices.

The parabola on Chart 19-9 is seen to be a much better fit than the straight line. That is, it follows the data more closely and roughly bisects most of the cycles in the period for which it was fitted.⁸ On the other hand, the shape of the parabola might be influenced so greatly by cyclical or irregular fluctuations that it may not be a satisfactory description of trend even if it fits the data much better than the straight line does. In particular, the parabola is dangerous for use in forecasting, as it tends to become unreasonably steep (or to turn down, if c is negative) when projected far into the future.

Third-degree polynomial trends of the form $Y_c = a + bX + cX^2 + dX^3$ and curves with still higher powers of X may also be fitted by the method of least squares, but these curves involve excessive labor and produce wavelike forms inconsistent with the concept of secular growth as a smooth curve. Therefore, these curves are seldom used for this purpose.

Logarithmic Straight Line. A straight line drawn on a ratio chart (sometimes called an exponential or compound-interest curve) is often more useful for trend analysis than either the arithmetic straight line or parabola described above. Many younger industries tend to expand at a constant percent rate of growth rather than at a constant amount of growth per year which appears as a straight line on an arithmetic chart. Furthermore, the arithmetic straight line is often illogical in that the constant amount of growth each year is independent of the size of the industry itself. Finally, the slopes of logarithmic straight lines show average *percent* rates of growth, and so they are comparable for series of different units or widely different size, whereas the slopes of trend lines on arithmetic scales are not comparable in such cases.

⁸ The goodness of fit could also be compared mathematically by computing the sum of the squared deviations $\Sigma(Y - Y_c)^2$ from each trend curve and dividing by $(N - k)$, where k is the number of constants (a, b, c) in the trend equation—i.e., two in a straight line and three in a parabola. The trend with the smaller value of $\Sigma(Y - Y_c)^2 \div (N - k)$ is the better fit by this criterion. See Mordecai Ezekiel and Karl A. Fox, *Methods of Correlation and Regression Analysis* (3d ed.; New York: John Wiley, 1959), Chap. 7, for a further discussion of this measure of goodness of fit.

Even if the rate of growth tends to diminish over a long period, the logarithmic straight line can be used to average the rate over some shorter interval, such as a decade, when the rate of change may be nearly constant.

Measurements of this type . . . possess the advantages of simplicity and ease of calculation. They lend themselves readily, moreover, to comparison and combination, since they are expressed in percent form. . . . This method yields, for each series, a single measurement which summarizes the direction and degree of change of that series during a stated period and which is directly comparable with similar measures derived from other series, regardless of the units of measurement in which the various series may have been expressed and of the magnitude of the figures in the various series.⁹

A logarithmic straight line may be fitted either graphically or by the method of least squares. The graphic method was applied to Sears, Roebuck sales earlier in the chapter, for the first thirty years of its department-store expansion period, 1926–1956. However, because of the retardation in the rate of growth after World War II, it appeared desirable to fit separate trends to the periods before and after the war. A trend is fitted by least squares below, therefore, to Sears, Roebuck sales in the postwar period 1947–1965.

In the *method of least squares*, look up the *logarithms* of the sales, then fit the equation $\log Y_c = a + bx$ exactly as in the least-squares solution for the arithmetic straight line, using $\log Y$ in place of Y .

In Table 19–3, the years (x) are listed in column 2 with the origin centered in 1956, sales are shown in column 3 in billions rather than millions to simplify the logarithms, the logarithms of the sales ($\log Y$) appear in column 4, and the product for each year ($x \log Y$) appears in column 5. Columns 4 and 5 are then totaled, and Σx^2 is found from Appendix K. To determine a and b (which are both logarithms in this equation),

$$a = \frac{\Sigma \log Y}{N} = \frac{10.7672}{19} = 0.5667$$

$$b = \frac{\Sigma x \log Y}{\Sigma x^2} = \frac{12.1416}{570} = 0.02130$$

The trend equation is therefore

$$\log Y_c = 0.5667 + 0.02130x \quad (\text{origin 1956})$$

⁹ Frederick C. Mills, *Economic Tendencies in the United States* (New York: National Bureau of Economic Research, 1932), p. 48.

Table 19-3

LOGARITHMIC STRAIGHT LINE FITTED BY LEAST SQUARES
TO SEARS, ROEBUCK DEFLATED NET SALES, 1947-1965

Year (1)	x (2)	Deflated Sales* (Billions) Y (3)	$\log Y$ (4)	$x \log Y$ (5)	$\log Y_c$ (6)	Trend Y_c (7)	Adjust- ment for Trend Y/Y_c (Percent) (8)
1947 . . .	-9	2.400	0.3802	-3.4218	0.3750	2.371	101.2
1948 . . .	-8	2.589	0.4131	-3.3048	0.3963	2.491	103.9
1949 . . .	-7	2.496	0.3972	-2.7804	0.4176	2.616	95.4
1950 . . .	-6	2.945	0.4691	-2.8146	0.4389	2.747	107.2
1951 . . .	-5	2.794	0.4462	-2.2310	0.4602	2.885	96.8
1952 . . .	-4	3.116	0.4936	-1.9744	0.4815	3.030	102.8
1953 . . .	-3	3.186	0.5032	-1.5096	0.5028	3.183	100.1
1954 . . .	-2	3.195	0.5045	-1.0090	0.5241	3.343	95.6
1955 . . .	-1	3.598	0.5561	-0.5561	0.5454	3.511	102.5
1956 . . .	0	3.828	0.5830	0	0.5667	3.687	103.8
1957 . . .	1	3.803	0.5801	0.5801	0.5880	3.873	98.2
1958 . . .	2	3.917	0.5930	1.1860	0.6093	4.067	96.3
1959 . . .	3	4.200	0.6232	1.8696	0.6306	4.272	98.3
1960 . . .	4	4.253	0.6287	2.5148	0.6519	4.486	94.8
1961 . . .	5	4.368	0.6403	3.2015	0.6732	4.712	92.7
1962 . . .	6	4.702	0.6723	4.0338	0.6945	4.949	95.0
1963 . . .	7	5.194	0.7155	5.0085	0.7158	5.198	99.9
1964 . . .	8	5.786	0.7624	6.0992	0.7371	5.459	106.0
1965 . . .	9	6.390	0.8055	7.2498	0.7584	5.733	111.5
Totals	0		10.7672	12.1416			

* Sales in billions of 1965 dollars, years beginning February 1, from Table 19-1.

To graph the trend on a ratio chart, plot any two widely separated points, using natural values of Y_c , and draw a straight line through them, as in Chart 19-10.

In 1947, $x = -9$,

$$\log Y_c = 0.5667 - 0.1917 = 0.3750 \quad \text{so } Y_c = 2.371$$

In 1965, $x = +9$,

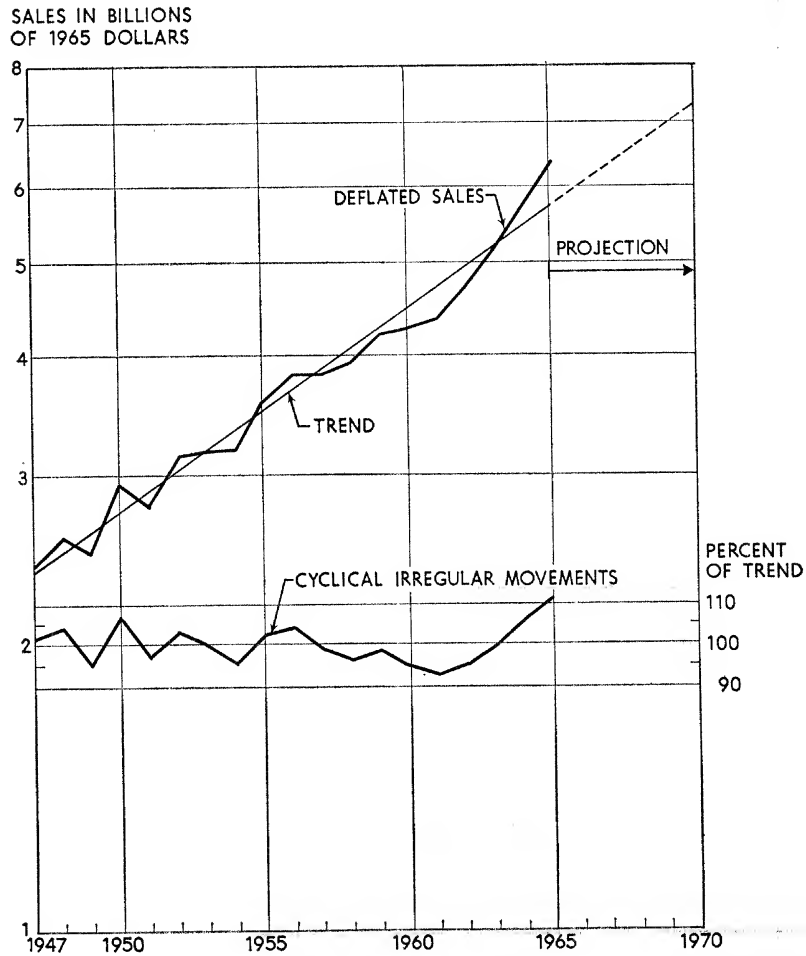
$$\log Y_c = 0.5667 + 0.1917 = 0.7584 \quad \text{so } Y_c = 5.733$$

As a forecast for 1970, $x = 14$, $\log Y_c = 0.8649$, and the trend forecast Y_c is 7.326 billion dollars. The *slope* of the least-squares trend line is the logarithm b . This means that the ratio of each year's trend value to the preceding year's is antilog b , or 1.050. The average rate of growth for 1947-1965 is then $1.050 - 1 = 0.050$ or 5.0 percent.¹⁰

¹⁰ The average rate of growth can be computed without using logarithms by the short-cut "method of moments," as follows: Take the first year as the origin (i.e., $X = 0$ in 1947), then compute XY for each year (where Y equals sales), and sum the Y and XY

Chart 19-10

LOGARITHMIC STRAIGHT LINE FITTED BY LEAST SQUARES
TO SEARS, ROEBUCK DEFLATED SALES, 1947-1965,
PROJECTED TO 1970



columns. Now compute the "mean value," $M = \Sigma XY / \Sigma Y$ and look up M and n (the number of years) in the Mean Value Table of James W. Glover's *Tables of Applied Mathematics* (Ann Arbor, Michigan: George Wahr, 1930), p. 471 f., to find the slope r . The a value may also be computed as described in Glover, p. 470.

This method minimizes the absolute deviations about the trend line rather than the logarithmic deviations, and so it gives more weight to larger values. The results of the method of moments and the logarithmic method do not differ appreciably, however, unless there are cyclical extremes at either end of the series. For further discussion, see Mills, *Economic Tendencies*, pp. 46-49; and Arthur F. Burns, *Production Trends in the United States since 1870* (New York: National Bureau of Economic Research, 1934), pp. 42-44.

This compares with the 7 percent growth rate determined graphically for the 1926–1956 period.

The trend may be eliminated, if desired, by computing and plotting Y/Y_0 , or antilog ($\log Y - \log Y_0$), for each year. The computations are shown in Table 19–3, columns 6 to 8. The resulting curve resembles the graphically adjusted curve at the bottom of Chart 19–6, except that the trend base is the logarithmic straight line rather than the growth curve.

The parabola and logarithmic straight line appear to fit the trend of Sears, Roebuck sales about equally well over the period 1947–1965. The latter, though, is generally preferable to the parabola because it is simpler and more rational in expressing growth as a constant percent per year, rather than as an arithmetic function of both time (x) and the square of time (x^2)!

The graphic and least-squares methods of fitting a logarithmic straight line give nearly the same results. The graphic method is recommended for quick, approximate results, and as a check on other methods, while the least-squares method is preferable for detailed, objective study, where computational assistance is available. The logarithmic least-squares method has the same merits and limitations as the arithmetic least-squares method described earlier in the chapter, except that the logarithmic straight line is more likely to be distorted by extreme *low* values than by extreme high values.

In summary, the trend analysis on Chart 19–10 shows that (1) Sears, Roebuck real sales have increased at an average rate of 5.0 percent per year from 1947 to 1965; (2) there is no recent evidence that the rate of growth is slowing down (though the average postwar rate is below the average prewar rate); (3) deviations from the trend “normal” have not exceeded $7\frac{1}{2}$ percent, except in 1965; (4) real sales can be projected over the next few years at an increase of 5.0 percent per year if the forces that made for past growth can be expected to persist.

The logarithmic straight-line projection gives a 1970 forecast of \$7.326 billion at 1965 prices, as noted above. However, if the cyclical-irregular prosperity level of 1965 (111.5 percent of trend in Table 19–3, column 8) is expected to continue unchanged, the forecast should be 7.326×111.5 percent = \$8.168 billion at 1965 prices. Finally, if a forecast in current dollars is needed, prices must also be projected. Thus, if we predicted an increase of 1 percent per year in Sears, Roebuck prices, the compounded increase would be 5.1 percent in the five years 1965–1970, and the 1970 forecast would be 8.168×1.051 = \$8.585 billion at current prices. However, this last

step is usually omitted because of the difficulties of forecasting price changes, and forecasts are usually expressed in terms of constant dollars.

The actual forecast of the cyclical-irregular element (which almost certainly will *not* remain unchanged at 111.5 percent of trend) requires the analysis of prospective changes in population and its age composition;¹¹ the correlation of sales with disposable personal income and other economic factors (as described in Chapter 24), together with available forecasts of the latter;¹² changes in consumer preferences; and the company's own expansion policy. Trend analysis is, of course, only the first step in long-range forecasting; the trend projection must be modified by a thorough study of all pertinent economic factors.¹³

SUMMARY

An understanding of the nature and causes of business fluctuations is essential in a dynamic economy. These fluctuations may best be understood by analyzing economic time series into their principal components—secular trend, seasonal variations, cyclical fluctuations, and irregular movements.

The trend and seasonal components are measured directly, while cyclical-irregular movements are usually treated as a residual in combined form.

Secular trend is the gradual long-term increase or decrease in a series resulting from such basic factors as the growth of population, technology, and productivity. This development can be represented by a smooth trend curve fitted to the plotted data. Different series vary greatly in the shape and steepness of their trends, as well as in the variations of the data from the trend curve. Young industries and total production tend to grow at a constant percentage rate. The rate of growth is often retarded as an industry matures, following the "law of growth" principle, and eventually tends to level off or even turn down.

Secular trend may be measured for three purposes: (1) the study of past trends, (2) long-term forecasting, and (3) the elimination of trend to isolate cycles. The period of years selected for trend analysis should be as long as possible in order to minimize short-term disturbances; it should be broken at points of abrupt change; and it should begin and end at the same stage of the business cycle.

¹¹ See U.S. Bureau of the Census, *Current Population Reports, Population Estimates*, Series P-25, Nos. 326, 329 (1966), *et seq.* for projections to 1985.

¹² See Economic Index and Surveys, *Predicasts* (quarterly) for forecasts of disposable personal income, other GNP components, and many industry figures.

¹³ See W. F. Butler and R. A. Kavesh, *How Business Economists Forecast* (Englewood Cliffs, New Jersey: Prentice-Hall, 1966); and H. D. Wolfe, *Business Forecasting Methods* (New York: Holt, Rinehart & Winston, 1966) for methodology.

Price deflation is the process of dividing a dollar value series by a pertinent price index in order to reveal physical volume changes, expressed in "constant dollars." An appropriate price index may be compiled from segments of existing indexes, properly weighted, as in the Sears, Roebuck example. Price deflation is particularly necessary in times of wide price changes, since the "real" changes in output may differ drastically from the reported dollar figures.

Trend may be measured by any of three methods: (1) a graphic "freehand" fit, (2) selected points, and (3) least squares. Annual data are usually used—preferably plotted on a ratio chart.

1. To fit a trend curve by the graphic method, draw it with a transparent ruler or French curve so as to equalize the areas or vertical deviations above and below each major segment of the curve. Averages of groups of years may be plotted as aids in locating the trend. The average growth rate of a logarithmic straight line can be read off the percent scale on the chart. To eliminate trend, lay off the vertical deviations from the trend line about a horizontal line on the ratio chart and label the scale "Percent of Trend."

Graphic methods are quick, flexible, and afford a continuous picture of successive steps, while mathematical methods are more objective and often more accurate; the latter can be performed by clerical labor or by electronic computers, and the results can be expressed in concise form. The two methods may be combined for optimum effectiveness.

2. The method of selected points is used in fitting "growth" curves. Growth curves of the logistic or Gompertz type represent the rational tendency of many industries and populations to grow at a declining percent rate as they mature. A curve of this type can be drawn graphically by using a transparent French curve on a ratio chart. It may also be fitted by selecting three typical points or cycle averages at equal intervals of time and computing the values of three constants in the appropriate equation, or by using a nomograph and special grid as a short-cut. Although subjective in nature, these curves are widely used both in the study of past trends and in forecasting.

3. The method of least squares fits a mathematical curve to the data such that the total of the squared deviations from the curve is less than that for any similar curve. The plus and minus deviations themselves total zero. This method is objective and reasonably accurate, provided the data follow the equation type chosen and are not too erratic. Unfortunately, however, the optimum conditions for the least-squares method do not occur in time series.

To fit a straight line by least squares, center the X origin at the middle year; set up a table of x , Y , xY , and x^2 , and substitute the

column totals in the given equations to find a and b in the equation $Y_c = a + bx$. To eliminate trend and isolate cyclical-irregular movements, compute and plot Y/Y_c for each year.

To fit a parabola, add columns for x^2Y and x^4 to the foregoing and substitute the totals in three equations to find a , b , and c in the equation $Y_c = a + bx + cx^2$. This is usually a better fit than a straight line, although it may be unduly affected by cyclical or irregular extremes.

The logarithmic straight line is superior to the other two in describing a rational growth tendency of young industries and in comparing relative rates of change. It may be drawn graphically as a straight line on a ratio chart or computed by the method of least squares. The least-squares procedure is the same as in the arithmetic straight line, except that $\log Y$ is used in place of Y . The projection of this function is often a reasonable first step in making medium-range forecasts for perhaps five or ten years in the future.

PROBLEMS

1. a) If you were an economist with the Eastman Kodak Company, manufacturers of camera and film (or other selected company), what would be the principal purpose of separating the company's monthly dollar sales into its component fluctuations? Give reasons to support your opinion.
- b) Briefly describe the causes of the four major components of this particular time series.
- c) Plot the company's annual sales for the past 15 or 20 years or trace them from an available chart.
- d) Describe the trend characteristics of this series: Is the trend a straight line, concave upward, or concave downward? What does this mean in terms of growth? Is the growth steady or erratic?
2. Select in the *Survey of Current Business* a price index that might be appropriate for deflating the gross revenues of each of the following:
 - a) A manufacturer of drugs and pharmaceuticals.
 - b) A Cleveland building contractor.
 - c) A clothing store.
 - d) A grocery supermarket.
- 3 and 4. Given the following data:

Year	Disposable Personal Income (Billions)	Average Hourly Earnings—Manufacturing Production Workers	Consumer Price Index (1957–1959 = 100)
1940	\$ 75.7	\$0.655	48.8
1945	150.2	1.016	62.7
1950	206.9	1.440	83.8
1955	275.3	1.86	93.3
1960	350.0	2.26	103.1
1965	465.3	2.61	109.9

SOURCE: *Survey of Current Business* (May 1966) and *Supplement, Business Statistics*, 1965.

3. *a)* Deflate disposable personal income by the consumer price index and list the results.
b) Plot actual and deflated income on a small chart.
c) Explain the significance of the deflated data and compare the trends of the two curves.
4. As a labor union economist, you wish to prepare a report summarizing the changes in real hourly earnings in manufacturing industries from 1940 to 1965, by five-year intervals. Besides eliminating changes in living costs, you feel that the results would be most meaningful if expressed in dollars of 1965 purchasing power, since recent price levels are most easily remembered. Using the above figures:
 - a)* Compute real hourly earnings in 1965 dollars.
 - b)* Compare the 1940–1965 percent increase in average hourly earnings with that in the real purchasing power of these earnings.
 - c)* To buy the same amount of goods and services that the 1965 worker could earn in one hour, how many hours would his father have had to work in 1940?
5. *a)* Under what conditions is it valid to forecast by extrapolating a trend curve fitted to past data? Discuss briefly.
b) Why may the particular purpose of measuring trend affect the choice of a trend curve?
c) What factors determine the period of years used in fitting a secular trend curve to an industry's sales?
d) Describe the use of group averages in trend fitting.
e) What is the one chief advantage of mathematical methods and of graphic methods, respectively, in trend analysis? Justify your selection.
6. *a)* Explain the "law of growth" principle implicit in the use of growth curves.
b) Describe briefly one method of fitting a growth curve.
c) What is the logical justification, if any, of fitting and projecting such a curve as a twenty-year forecast of aluminum production (Chart 19–2)?
7. As part of a planning study for General Foods Corporation, you are asked to analyze and project the growth trend in the output of manufactured food products, as measured by the Federal Reserve Index of Production in Food Manufactures (1957–1959 = 100), shown below. While this index has been compiled since 1947, it is felt that in such a stable industry as food products, even the short period 1957–1965 might provide a reliable picture of current trends.

Year	Index	Year	Index
1957	96.9	1962	113.8
1958	99.4	1963	116.8
1959	103.8	1964	120.1
1960	106.9	1965	122.4
1961	110.6		

SOURCE: *Federal Reserve Bulletin or Survey of Current Business.*

- a) Plot this series on an arithmetic chart. Since the growth is roughly linear, fit a straight-line trend by the method of least squares. To save labor, express the index as a deviation from 100 (i.e., subtract 100).
- b) State the average annual growth from 1957 to 1965 (give the unit). Compute Y/Y_c for 1965 to find the cyclical-irregular component, or the value "adjusted for trend," in that year (give unit).
- c) Plot the trend line on the chart and extend it beyond 1965 to the latest year for which the index is available. Multiply the projected trend value by the cyclical-irregular component for 1965 (assuming that this factor continues unchanged) to obtain a forecast. Find the actual food manufactures index for this year and give the percent error of the forecast. Explain the probable causes of this error.

8 to 11. As an economist in the chemical industry, you wish to analyze and forecast the postwar growth in chlorine production, shown here in millions of short tons:

Year	Chlorine Production	Year	Chlorine Production	Year	Chlorine Production
1947	1.45	1954	2.90	1961	4.60
1948	1.64	1955	3.42	1962	5.14
1949	1.77	1956	3.80	1963	5.46
1950	2.08	1957	3.95	1964	5.94
1951	2.52	1958	3.60	1965	6.44
1952	2.61	1959	4.35		
1953	2.80	1960	4.64		

SOURCE: *Survey of Current Business* (May 1966) and *Supplement, Business Statistics*, 1965.

- a) Plot these figures on a one-cycle ratio chart, with the time scale extended to date. Use the proper title, scale captions, and labels for curves.
- b) Draw a smooth "growth" curve (slightly concave downward) through the data by inspection, and adjust it so that the vertical deviations above and below are about equal for each major segment (the deviations may be cumulated on a paper strip). Extend the curve to date as a forecast on the assumption that some retardation in growth rate will occur after 1965.
- c) Draw a logarithmic straight line through the data beginning in 1951 by inspection, and extend it to date on the more optimistic assumption that the average 1951-1965 rate of growth will continue unchanged. Find the average annual rate of growth graphically and state it as a percent.
- d) Forecast chlorine production beyond 1965, using (1) the trend in *b* or *c* that appears the more reasonable and (2) a cyclical-irregular adjustment (either as a percent of trend or as a vertical distance laid off on the chart) based on 1965 production relative to trend, modified by your best judgment. Explain the reasons for your procedure.
- e) Look up actual chlorine production in the years following 1965 in the *Survey of Current Business*, plot it on the chart, and note the percent error in your forecast for the latest year. What is the probable reason for this error?

9. *a)* Eliminate the trend in Problem 8 graphically (using the trend curve you prefer), and plot the cyclical-irregular relatives in the lower part of the chart.
- b)* Describe the cyclical timing and amplitude of chlorine production, and the principal irregular forces at work, during the postwar period.
10. *a)* Plot chlorine production for 1951–1965 on an arithmetic chart.
- b)* Fit either a straight line or parabola by least squares, depending on which appears to be a better fit.
- c)* Using this trend, project future chlorine production and compare with actual results, as described in Problem 8 (*d*) and (*e*) above.
11. *a)* Fit a logarithmic straight line by least squares to chlorine production, 1951–1965, and extend it beyond 1965.
- b)* Find the average annual rate of growth, using logarithms.
- c)* Compare the goodness of fit of the logarithmic straight line fitted graphically with that fitted by least squares.

Problems 12 to 15 may be assigned either for full-length analysis, as given, or as short illustrative exercises covering only the seven years beginning 1959.

- 12 to 15. The annual production of electricity by electric utilities in the United States from 1947 to 1965 was as follows (in billions of kilowatt-hours):

Year	Electricity Production	Year	Electricity Production	Year	Electricity Production
1947	256	1954	472	1961	792
1948	283	1955	547	1962	852
1949	291	1956	601	1963	914
1950	329	1957	632	1964	984
1951	371	1958	645	1965	1,055
1952	399	1959	710		
1953	443	1960	753		

SOURCE: *Survey of Current Business*.

12. *a)* Plot these figures on a one-cycle ratio chart, with the vertical scale beginning at 200 billion kilowatt-hours and the horizontal scale extended beyond 1965, up to date.
- b)* Draw a smooth freehand trend line or curve through the data, and project it several years beyond 1965, plotting group averages as guides and equalizing the deviations above and below the trend as described in the text.
- c)* Describe the nature of growth in this industry. What has been the average annual percent rate of growth since 1959? (Show on the chart how this value was obtained.)
13. Plot electricity production on an arithmetic chart, with the time scale extended beyond 1965, and compute an arithmetic straight line by the method of least squares. Show computations and trend equation. Plot this curve on the arithmetic chart and project it into the future.

14. *a)* Fit a logarithmic straight line to the same data by least squares, plot it on the ratio chart, and extend it beyond 1965.
b) How does the least-squares criterion of goodness of fit differ in its application to the arithmetic straight line and the logarithmic straight line?
c) Explain the meaning of the constants a and b in each of these equations.
15. *a)* Compare the goodness of fit of the freehand trend, the arithmetic straight line, and the logarithmic straight line in describing the growth of electricity production.
b) Which of these three curves is the most logical for use in forecasting? Why?
c) Find comparable up-to-date figures on electricity production and plot them on your two charts. What are the percent errors in your forecasts for the latest year? What factors might explain these errors?
16. General Electric Company sales began a marked upward climb beginning in 1959. As a market analyst with this company, you are asked by your department head to determine the average yearly rate of growth in the physical volume of sales from 1959 to 1965. This may be done by computing the slope of a logarithmic straight line (not by averaging year-to-year changes, which have different bases), as fitted to deflated sales. The 1965 Annual Report gives these sales (in billions of dollars) and price indexes of General Electric products (1957-1959 = 100):

	1959	1960	1961	1962	1963	1964	1965
Sales	4.47	4.38	4.67	4.99	5.18	5.32	6.21
Price index	101	97	92	90	87	87	87

- a)* Express sales in terms of 1965 dollars.
b) Fit a logarithmic straight line by least squares to the deflated sales.
c) Find the average annual percent rate of growth.
d) Apply this rate of increase to 1965 deflated sales and assume a price index of 89 in 1966 to estimate 1966 actual sales. Compare this with reported sales.

SELECTED READINGS

Readings for this Chapter have been included in the list which appears on page 549.

20. SEASONAL VARIATION

OF THE PRINCIPAL types of fluctuations in economic activities, trend analysis was discussed in Chapter 19. In this chapter the purposes and principal methods of measuring seasonal variation are surveyed. Cyclical and irregular fluctuations will be considered in Chapter 21.

In trend analysis, annual data are usually used. For the study of shorter-term seasonal and cyclical movements, however, quarterly, monthly, or weekly data are needed. Monthly figures are most common.

NATURE OF SEASONALITY

Seasonal variations are of two kinds: (1) those resulting from natural forces and (2) those resulting from man-made conventions. For example, in the northern United States and Canada, construction work is greatly curtailed during the winter season. Hence, data concerning road construction, building activity, and the like have seasonal variations that are directly related to the weather. On the other hand, department store sales expand before Easter and Christmas, a circumstance related to man-made festivals rather than to the weather.

Seasonal variations affect nearly all economic activities. The impact of seasonal influences is likely to be greatest at the point of origin and the point of consumption and less in the intervening manufacturing process. The cotton crop, for example, is seasonal, and so are retail sales of cotton goods (in a different pattern), but textile mills manage to operate at a more stable rate by manufacturing for stock in the slack seasons. In some industries, however, only the supply is markedly seasonal (e.g., wheat versus bread) or the demand (consumer durable goods) or the fabrication process itself (building construction). Inventories in general are more seasonal, and prices less seasonal, than pro-

duction or sales. The typical seasonal pattern includes either one peak and trough per year, as in building construction, or else peaks in both spring and fall and troughs in midwinter and midsummer, as in retail trade generally.

The latter pattern is illustrated by the monthly sales of Sears, Roebuck shown in Chart 20-2. The year starts with the midwinter slump, followed by a brisk spring trade, a June dip, a fall pickup, and a big Christmas rush. Accurate measures of seasonal behavior by products are invaluable to the management of such a firm in planning purchasing, inventory control, and selling programs.

Two important features of the seasonal rhythm should be noted: (1) it recurs year after year with a fixed period and (2) the increases and decreases of sales occur at about the same time and in about the same proportion each year.¹ The seasonal rhythm therefore has a fixed period and a fairly regular amplitude, whereas the cyclical rhythm is variable in both respects. Seasonal movements, consequently, may be measured and projected into the future much more accurately than cycles.

Calendar Variation

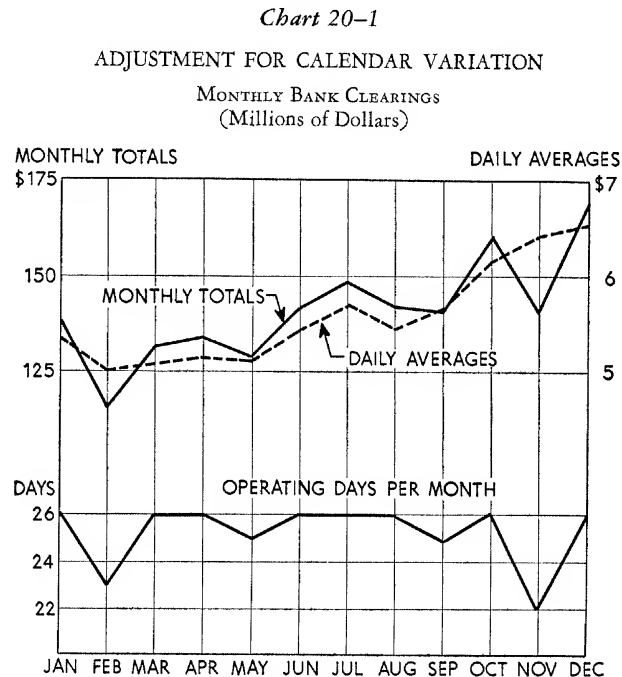
One cause of "seasonal" disturbances in monthly and weekly data is neither the weather nor customs but the eccentricity of the calendar itself. The months not only vary from 28 to 31 days in length, but some have four Saturdays and Sundays, others have five. Some also have one or several holidays, others have none. Further, certain series of data arise from activities which operate five days a week, others 5½, 6, or even 7 days. All these factors cause spurious movements in monthly data which cannot be entirely eliminated by seasonal adjustment.

It is usually desirable, therefore, to eliminate the effect of calendar variation as a preliminary step before measuring regular seasonal movements. The method of adjusting for calendar variation is to divide each monthly total by the number of operating days in that month to reduce it to a uniform average daily basis. The general rule is to count the number of days that the particular activity was carried on during the month. In some cases this will mean all of the days in the month; in others, Sundays or Saturdays, Sundays, and holidays will be excluded. If one day in the week is unusually light or heavy in volume, it may be weighted accordingly. Thus, the Federal Reserve Board weights Sunday

¹Two notable exceptions occur because (1) the date of Easter varies and (2) automobile production and sales are affected by the variable dates of offering new models. These irregularities require special corrections in seasonal measurement.

as $1\frac{1}{2}$ days in adjusting monthly newspaper output—a component of the Industrial Production Index. Different holidays are also observed in the several fields of business activity and in different areas.²

Chart 20-1 shows the effect of calendar adjustment on a city's monthly bank clearings in a leap year when banks were closed Sundays and eleven holidays. The monthly totals are divided by the number of operating days per month (bottom curve) to yield the daily averages (dashed line, right scale). It is evident that most of the month-to-month



fluctuations in total clearings—particularly the dips in February and November—were due merely to the erratic calendar and not to any significant change in banking activity.

The method of reducing to a daily average basis should be used only for quantities that cumulate during the month, such as bank clearings, production, or sales. These series all add up to larger amounts in long months than in short months. On the other hand, series such as bank deposits, prices, employment, or other "point data" should not be re-

² For a list of weekly working days in the principal manufacturing, mining, and utility industries, see Federal Reserve Board, *Industrial Production: 1957-59 Base* (1962), pp. S-4 to S-19. See also A. Young, *Estimating Trading-Day Variation in Monthly Economic Time Series* (Technical Paper No. 12) (Washington, D.C.: U.S. Bureau of the Census, 1964).

duced to an average daily basis, because they do not cumulate or build up to larger values in longer months. Yearly and quarterly data in general are not adjusted for the calendar either, since the irregularity is negligible in these longer periods.

In the case of weekly data the number of weekdays is constant, and only holidays cause irregularities. These may be corrected by (1) adjusting weeks containing holidays to a full-time basis (e.g., adding one fourth to the figure for a four-day week to make it comparable with data for five-day weeks) or (2) plotting curves for one year over the other on a tier chart so that weeks containing a given holiday are lined up vertically for direct comparability in different years, as in Chart 20-6.

When data are to be adjusted for seasonal variation, as described below, the calendar adjustment may sometimes be omitted, since the seasonal correction eliminates the difference between the *average* number of operating days in January and those in February. However, it does not smooth out the differences in operating days between one January and the next. Thus, if one January had 26 days and the next had 27 days, and we divided the two January totals by the same seasonal index, the adjusted data would still show a spurious difference due to the calendar.

Other Rhythms

Many economic activities exhibit rhythmic movements having a shorter period than seasonal variations. Quarterly dividend and income tax payments and monthly payrolls cause regular fluctuations in the flow of funds through banks and in consumers' expenditures. Weekly rhythms may be illustrated by the sales in a department store. Monday is apt to be light, except after a long holiday weekend; then trade builds up gradually during the week to a peak on Saturday. The average sales on a number of Mondays may be compared with the averages for other weekdays (with separate norms for days before and after holidays) and a normal pattern of weekly variation worked out to aid in the timing of purchasing, advertising, and hiring of extra help.

Daily rhythms occur in such data as the hourly number of messages crossing a telephone switchboard, the hourly number of riders on buses, or the hourly use of electric power. These and many similar series have such regular fluctuations that engineers use them to determine the amount of equipment to be kept in service each hour of the day and night.

The rhythms having a shorter period than the seasonal, therefore, may be worth analyzing as an aid to short-term programming. Since they

do not require the use of statistical techniques beyond averages, however, no further attention will be given to them here.

PURPOSES OF MEASURING SEASONALITY

There are three principal purposes of measuring seasonal movements: (1) to analyze past seasonal behavior, (2) to predict seasonal movements as an aid in short-term planning, and (3) to eliminate seasonality in order to reveal cyclical movements.

1. Measures of typical seasonal behavior in production, sales, inventories, and prices are indispensable in understanding the characteristic fluctuations of a business during the year and in gauging the significance of current figures. Seasonal indexes serve to answer such questions as: Was the decline in sales last month more or less than the usual seasonal amount? How much does the price of a given product usually decline between July and August? What is the normal variation in inventories from month to month?

2. Seasonal measures are also useful in planning operations over the next year or two. Every successful business concern operates on a budget, in which the coming year's income and expense items are estimated, and later checked against actual results. By means of seasonal indexes, next year's budget items may be allocated by months. Seasonal indexes are also particularly useful in scheduling purchases, personnel requirements, seasonal financing, and selling and advertising programs. Seasonal movements, like cycles, are wasteful because the men and equipment needed in the peak season are idle in the slack season. An accurate knowledge of seasonal behavior is an aid in mitigating and ironing out seasonal movements through business policy. This may be done by introducing diversified products having different seasonal peaks, accumulating stocks in slack seasons in order to manufacture at a more regular rate, cutting prices in slack seasons, and advertising off-seasonal uses for products.

3. Perhaps the principal purpose of measuring seasonal variations is to get rid of them. Business cycles are of critical importance, but these cycles are frequently obscured by large seasonal movements. The latter must ordinarily be measured and eliminated to reveal the former. Many monthly statistical series in economic publications are "adjusted for seasonal variation" for this purpose. The *Survey of Current Business*, for example, lists the following data and many others on a seasonally adjusted or simply "adjusted" basis: gross national product, industrial production, business sales and inventories, manufacturers' orders, new

construction, retail sales, and employment. A knowledge of seasonal adjustment therefore is essential for the economic analyst.

METHODS OF MEASURING SEASONAL VARIATION

Seasonal variation has been defined as a rhythmic movement which recurs each year with about the same relative intensity. This movement may be summarized by a seasonal pattern which is assumed to be typical of any year of a series or which changes gradually from year to year. The pattern consists of twelve monthly indexes (or four quarterly indexes) whose average is 100 percent. The problem of measuring seasonal variation is then one of determining these indexes for a given series.

A great many methods have been advanced for computing seasonal indexes. Essentially, however, most refined methods arrive at a seasonal index for a given month by averaging its ratios to a trend-cycle base in several years (or fitting a trend curve to these ratios) to cancel out the nonseasonal factors.

In any method of measuring seasonality the series is first plotted on a chart to show the general nature of the seasonal pattern and to aid in further analysis. Unless a fairly pronounced and regular rhythm is apparent, seasonal measurement may not be worthwhile. A ratio scale must be used in the graphic method described below and is usually desirable in other methods as well, since seasonal movements in most economic data are more stable as percentages than in absolute amounts. Hence, seasonal indexes themselves are expressed as percentages.

The period of time covered should be at least six or seven years for series having a regular seasonal pattern, and longer for irregular data, in order to average out the peculiarities in individual years. The conditions in this period should approximate those expected in the future if the seasonal indexes are to be used for forward planning. The normal seasonal rhythm may be disrupted by wars, strikes, government edicts, severe depressions, and abrupt changes in business policy. Such erratic periods should be excluded, as far as possible. Sometimes the seasonal nature of a series will change gradually over the years. In this case a relatively long period of years should be used, as in trend analysis, and "changing" indexes of seasonal variation should be computed as described later in the chapter. Such analysis might well begin with the year 1947 or 1953 for many series, since the disruptions of World War II and the Korean War distorted normal seasonal behavior throughout the war periods.

Graphic Method

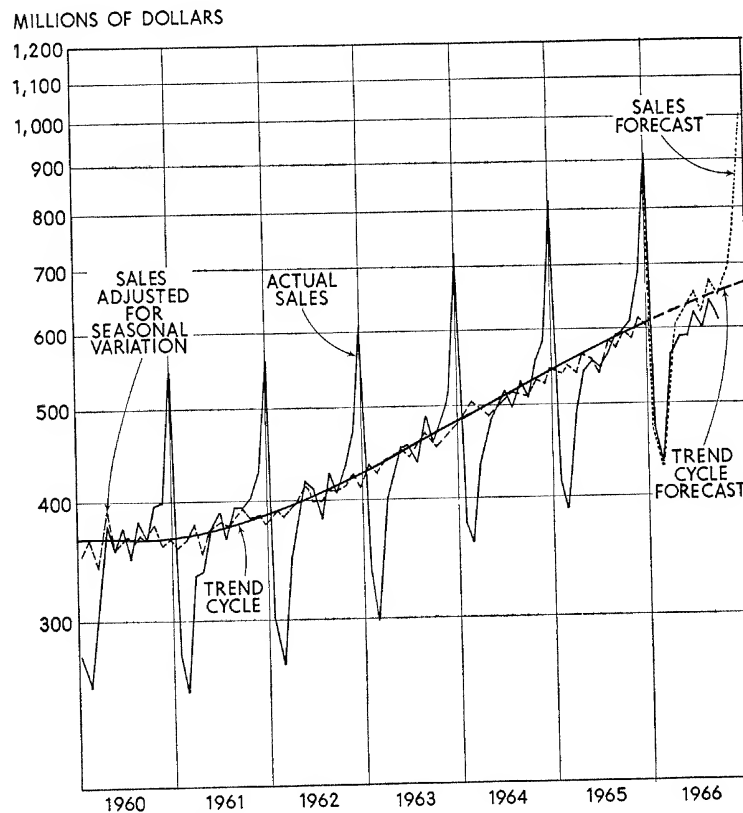
In the graphic short-cut method, most of the steps are performed directly on the chart. This technique will be applied to monthly sales of Sears, Roebuck from 1960 to 1965.³ The steps are:

1. Plot the data on a ratio chart, preferably with a one-cycle scale.

Chart 20-2

GRAPHIC SEASONAL METHOD

SEARS, ROEBUCK SALES, 1960-1966
Ratio Chart



SOURCE: Tables 20-1 and 20-2.

³ Sears, Roebuck and Co. sales have not been adjusted for calendar variation because the seasonal indexes themselves will reflect the difference in average length of months and correct for this in the adjusted data. Slight variations due to the varying number of weekdays between one January and the next etc. remain, and should be corrected by a separate calendar adjustment in a more refined study.

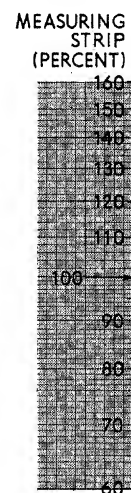
It is not necessary to deflate sales for price changes in seasonal analysis, since they have little effect on the seasonal rhythm and tend to cancel out in the averaging process.

The large scale makes measurements more accurate than on a two- or three-cycle paper, and the ratio scale permits measuring and averaging percentages on the graph. As shown in Chart 20-2, Sears, Roebuck sales have a pronounced seasonal rhythm, so that seasonal analysis is worthwhile.

2. Plot the annual average of monthly sales at the middle of each year (between June and July) and draw a freehand trend-cycle curve through these points (say, in red) by inspection. The curve should follow not only the trend but also cyclical and extended irregular movements such as those caused by war. A knowledge of economic conditions in this period will also help in locating the peaks and troughs of cycles.

Thus, the period 1960-1965 was marked by a recession from a peak in May 1960 to a trough in February 1961 and a continuous expansion thereafter.⁴ The fitting of this curve involves a subjective error, but part of the error is canceled in subsequent operations,⁵ and the curve can be altered later to improve the fit, if necessary, as described under "Revision for Greater Accuracy" below. The trend-cycle curve in Chart 20-2 is drawn horizontally from January 1960 through the recession trough of February 1961, since sales each month were at about the same level as a year ago, on the average. (Sales in 1959 are not reproduced here.) Thereafter, the trend-cycle curve is drawn with a French curve on a rising trend, passing through the annual averages, without any cyclical dips.

3. Take another sheet of one-cycle ratio graph paper, and lay off a percentage scale on its right margin, as illustrated, marking 100 percent with a red arrow opposite the number "5" printed on the graph paper, 120 percent opposite "6," 80 percent opposite "4," and the other numbers in the same proportion. Cut out a vertical *measuring strip* containing these values. Find the percentage of sales to the trend-cycle base for each month by placing the 100 percent red arrow of the measuring strip on the trend-cycle curve of the sales chart and reading off the value on the measuring strip opposite the plotted sales. (Check one or two of those measurements arithmetically by dividing sales by the trend-cycle value read from the chart.) Tabulate the percentages, as in Table 20-1. Dividing sales by the trend-cycle base



⁴ According to the National Bureau of Economic Research reference dates shown in Table 21-1.

⁵ The error cancels out either if the average level of the freehand curve is too high or too low (since the seasonal indexes are adjusted to average 100 percent) or if its positive and negative errors are equal (since the ratios for each month are averaged).

eliminates most of the influence of trend and cycle, so that the percentages reflect primarily the effect of seasonal and irregular movements. By averaging these percentages for a given month (step 4) the irregular factors tend to cancel out and the average itself reflects the seasonal influence alone.

4. Compute a modified mean of the percentages for each month in the different years, omitting the highest and lowest values as being unduly influenced by irregular factors such as strikes or a break in the stock market.

In Table 20-1 the highest figure and the lowest figure in each column are crossed out and the remaining four items are totaled and

Table 20-1
PERCENTAGES OF GRAPHIC TREND-CYCLE CURVE
AND COMPUTATION OF SEASONAL INDEXES

SEARS, ROEBUCK SALES, 1960-1965

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1960	74	70	82	103	97	103	96	104	100	108	109	149	
1961	75	69	91	91	101	104	96	104	103	105	111	145	
1962	78	69	88	96	104	102	94	103	98	104	112	144	
1963	79	69	89	97	101	101	96	105	97	101	110	149	
1964	78	75	88	94	98	102	96	101	97	105	110	150	
1965	76	70	86	97	98	96	97	101	101	103	114	149	
Total, middle four	307	278	351	384	398	408	384	412	396	417	443	592	
Modified mean	76.8	69.5	87.8	96.0	99.5	102.0	96.0	103.0	99.0	104.2	110.8	148.0	1,192.6
Seasonal index	77.3	70.0	88.3	96.6	100.1	102.6	96.6	103.6	99.6	104.9	111.5	148.9	1,200.0

divided by 4 to give the modified means shown in the next to the bottom row. These means are preliminary seasonal indexes. They should average 100 percent, or total 1,200 for 12 months, by definition. The total in Table 20-1, however, is 1,192.6, because extreme values have been dropped before averaging the rest.

5. Therefore, multiply each of the 12 modified means by the quotient of 1,200 over their total to yield the final *seasonal indexes*. Here, each mean is multiplied by $1,200/1,192.6$ and the resulting indexes are listed in the last row. They total 1,200 and hence average 100 percent.

The individual percentages and seasonal indexes in Table 20-1 are plotted in Chart 20-3, the seasonal indexes being connected by straight lines.

These indexes of seasonal variation provide a quantitative measure of typical seasonal behavior and a basis for future planning. The slumps in January, February, and July, the autumn rise, and the December peak

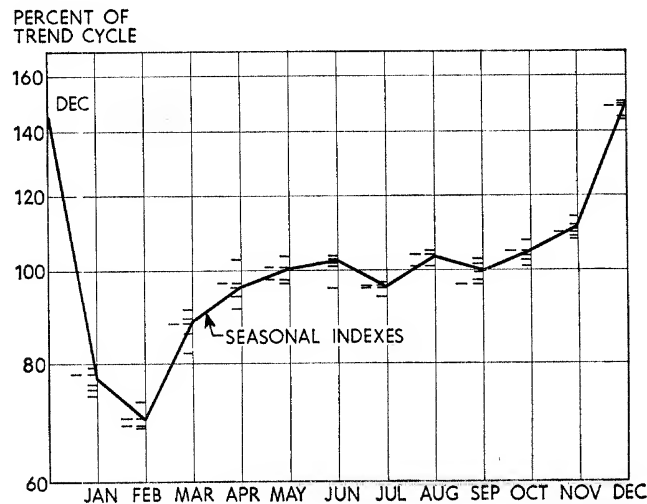
are clearly evident. The volume ranges from a low of 70 percent of the average month, in February, to more than double that volume, 149 percent, in December. The normal seasonal rise from November to December is 33 percent, that is, $(149 - 112)/112$ —the decline from December to January is 48 percent, and so on. (The seasonal indexes are rounded here since they are only accurate to the nearest percent.)

The irregularities in seasonal behavior are shown by the scatter of the percentages of trend-cycle for a given month in Chart 20-3. If the per-

Chart 20-3

SEASONAL INDEXES AND PERCENTAGES OF TREND-CYCLE—
GRAPHIC METHOD

SEARS, ROEBUCK SALES, 1960-1965



SOURCE: Table 20-1.

centages are closely bunched, it means that the seasonal standing of the month is regular from year to year and the seasonal index is reliable for use in forecasting. If all the scatters were centered about the 100 percent line, as in September, there would be no significant seasonality. In the case of Sears, Roebuck sales, however, the average seasonal movement shown by the displacement of the clusters away from the base line is unmistakable.

6. If it is desired to adjust the data to eliminate seasonal variation, mark the January seasonal index on the measuring strip, place this mark on each January sales point of Chart 20-2, and plot the adjusted value on the chart opposite the 100 percent arrow of the measuring strip. This has the effect of dividing actual sales by the seasonal index (e.g., for

January 1960, $271.3 \div 77.3 \text{ percent} = 351.0$). Do this for all months, raising the values for months with seasonal indexes below 100 and lowering those with indexes above 100. (The span between the seasonal index and 100 can be laid off on a blank sheet for convenience in adjusting different months.)

The adjusted sales for all months, drawn as a dashed line in Chart 20-2, reflect the trend, cycle, and irregular movements of the data, eliminating only the typical seasonal rhythm. This curve shows that Sears, Roebuck sales were depressed only slightly by the general business cycle decline from May 1960 and February 1961 and that in the long business expansion that followed Sears, Roebuck sales increased steadily. The month-to-month irregularities are due to calendar variation, the changing date of Easter, unusual weather conditions, special sales, and numerous unidentifiable causes. These irregularities can be smoothed out graphically or by a short-term moving average, as described in Chapter 21, to reveal the trend-cycle pattern of sales.

Revision for Greater Accuracy. The graphic method can be refined for more accurate results as follows: Draw a revised trend-cycle curve on the ratio chart so as to bisect the *seasonally adjusted* data, following the cyclical drift and ignoring only the month-to-month zigzag movements. The revised trend-cycle curve is shown in Chart 21-1. Then repeat steps 3 to 5 (and step 6 if the data are to be adjusted for seasonality), using the new curve. The revised trend-cycle curve is more sensitive to the cyclical positions of individual months than the original curve. Hence, the seasonal indexes are better. The correction in this case, however, does not seem to justify a revision. The same procedure can be used to improve the results of the 12-month moving-average method described below.⁶

Moving-Average Method

The moving-average method of measuring seasonal variation involves the same basic steps as the graphic method except that the steps are performed arithmetically. This method will be illustrated by the same Sears, Roebuck sales data as before. The steps are as follows:

1. Plot the series either on an arithmetic scale, for easier plotting, or on a ratio scale, to show seasonal swings of more uniform amplitude.
2. Compute a 12-month moving average to represent the trend-cycle

⁶ A method developed by the Federal Reserve Board combines the graphic and moving-average methods and adds other steps (fifteen in all) to refine the results, although at the cost of considerable additional labor. See H. C. Barton, Jr., "Adjustment for Seasonal Variation," *Federal Reserve Bulletin*, (June 1941), pp. 518-28.

base. This is simply a yearly average moved up a month at a time. A 12-month average includes both the high and low seasonal months during the year, and so the seasonal influences cancel out and the trend and cycle remain. The 12-month moving average is more objective than the freehand trend-cycle curve, although it tends to cut corners at cyclical turning points.⁷

To compute a 12-month moving average, first find the moving *total* as follows: Add the first 12 figures on an adding machine, list the total with the "subtotal" key on the tape, then add the next month and subtract the first month, list the subtotal again, and so on throughout the series. Check the last subtotal against an independent total of the last 12 months to verify all totals.

List each total in a table opposite the *seventh* of its 12 months.⁸ Then divide the totals by 12 to get the moving averages. This may be done most easily by entering the reciprocal of 12—0.083333—in a calculating machine and multiplying it successively by each of the totals without clearing the machine.⁹

In Table 20-2, Sears, Roebuck sales are listed from July 1959 to May 1966 to determine the moving averages for the six-year period January 1960 to December 1965, since they cannot be computed for the end months. The total for the first 12 months, July 1959–June 1960, is listed in column 3 opposite the seventh month, January 1960. Moving up a month, the next 12-month total for August 1959–July 1960 is computed as $4,329.3 + 349.6 - 343.9 = 4,335.0$ and listed opposite the seventh month, February 1960, and so on. These totals are

⁷ The 12-month moving average does not show the true trend-cycle position of its middle month but rather the average level of 12 adjoining months. Hence, it cannot reach the peaks, valleys, and extremities of a series; it errs in the direction of curvature in either trend or cycle, and distorts the 12 months centered on a point of abrupt change.

⁸ A 12-month total or average can be centered on either the sixth or seventh month, but the latter is a month more up to date. The exact center is midway between the two, so that sometimes two adjoining 12-month moving totals are themselves averaged in order to center exactly on a given month. Thus, a total of July 1959–June 1960 and August 1959–July 1960 would center precisely on January 1960. The steps are as follows: (1) Compute a 12-month moving total, listing the first item opposite the sixth month. (2) Compute a two-item moving total of these totals, entering the first item opposite the seventh month of the original data. (3) Divide by 24. This is the centered moving average. However, since the moving average is only a rough approximation of trend-cycle at best, this very minor refinement in timing does not appear to justify the considerable extra labor.

⁹ Twelve-month moving averages are used here to clarify the method, but the moving totals themselves can more easily be used in subsequent steps to save the labor of multiplying through by $\frac{1}{12}$, as follows: (1) Divide each month's sales by the moving total. The results will be just $\frac{1}{12}$ the percents of moving averages. (2) Compute the modified mean of these ratios for each month and total the 12 means. (3) Multiply each mean by 1,200 over this total to arrive at seasonal indexes identical with those in the text, the final multiplication factors being just 12 times those in the text method.

Table 20-2
COMPUTATION OF 12-MONTH MOVING AVERAGES

SEARS, ROEBUCK SALES, 1960-65

Month	Sales (Mil- lions)	12- Month Moving Total	12- Month Moving Average	Percent of Moving Average (Column 2 ÷ 4)	Month	Sales (Mil- lions)	12- Month Moving Total	12- Month Moving Average	Percent of Moving Average (Column 2 ÷ 4)
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1959:					1963:				
July	343.9	January	338.1	5,087.7	424.0	79.7
August	366.3	February	298.1	5,144.3	428.7	69.5
September	355.8	March	390.0	5,205.7	433.8	89.9
October	395.4	April	428.2	5,253.1	437.8	97.8
November	398.7	May	452.0	5,299.4	441.6	102.4
December	531.4	June	455.0	5,347.7	445.6	102.1
1960:					July	439.7	5,449.5	454.1	96.8
January	271.3	4,329.3	360.8	75.2	August	485.6	5,488.5	457.4	106.2
February	256.7	4,335.0	361.3	71.0	September	452.6	5,546.0	462.2	97.9
March	301.1	4,348.8	362.4	83.1	October	477.7	5,593.1	466.1	102.5
April	377.8	4,357.7	363.1	104.1	November	519.9	5,634.9	469.6	110.7
May	354.8	4,360.6	363.4	97.6	December	712.6	5,682.4	473.5	150.5
June	376.1	4,361.6	363.5	103.5	1964:				
July	349.6	4,376.8	364.7	95.9	January	377.1	5,747.1	478.9	78.7
August	380.1	4,381.7	365.1	104.1	February	355.6	5,804.8	483.7	73.5
September	364.7	4,378.3	364.9	99.9	March	437.1	5,848.5	487.4	89.7
October	398.3	4,411.7	367.6	108.4	April	470.0	5,905.2	492.1	95.5
November	399.7	4,371.5	364.3	109.7	May	499.5	5,984.2	498.7	100.2
December	546.6	4,391.3	365.9	149.4	June	519.7	6,051.2	504.3	103.1
1961:					July	497.4	6,153.9	512.8	97.0
January	276.2	4,405.4	367.1	75.2	August	529.3	6,192.5	516.0	102.6
February	253.3	4,417.8	368.2	68.8	September	509.3	6,224.8	518.7	98.2
March	334.5	4,433.0	369.4	90.6	October	556.7	6,263.3	521.9	106.7
April	337.6	4,461.2	371.8	90.8	November	586.9	6,336.1	528.0	111.2
May	374.6	4,463.9	372.0	100.7	December	815.3	6,395.8	533.0	153.0
June	390.2	4,494.0	374.5	104.2	1965:				
July	362.0	4,508.8	375.7	96.4	January	415.7	6,426.1	535.5	77.6
August	395.3	4,536.1	378.0	104.6	February	387.9	6,491.8	541.0	71.7
September	392.9	4,553.0	379.4	103.6	March	475.6	6,552.3	546.0	87.1
October	401.0	4,567.7	380.6	105.4	April	542.8	6,637.8	553.2	98.1
November	429.8	4,611.0	384.3	111.8	May	559.2	6,691.8	557.7	100.3
December	561.4	4,655.4	388.0	144.7	June	550.0	6,787.0	565.6	97.2
1962:					July	563.1	6,879.5	573.3	98.2
January	303.5	4,675.1	389.6	77.9	August	589.8	6,941.7	578.5	102.0
February	270.2	4,696.2	391.4	69.0	September	594.8	6,992.3	582.7	102.1
March	349.2	4,725.1	393.8	88.7	October	610.7	7,079.6	590.0	103.5
April	380.9	4,737.4	394.8	96.5	November	682.1	7,123.3	593.6	114.9
May	419.0	4,767.8	397.3	105.5	December	907.8	7,152.2	596.0	152.3
June	409.9	4,809.6	400.8	102.3	1966:				
July	383.1	4,859.0	404.9	94.6	January	477.9
August	424.2	4,893.6	407.8	104.0	February	438.5
September	405.2	4,921.5	410.1	98.8	March	562.9
October	431.4	4,962.3	413.5	104.3	April	586.5
November	471.6	5,009.6	417.5	113.0	May	588.1
December	610.8	5,042.6	420.2	145.4					

then multiplied by $\frac{1}{12} = 0.083333$ with a calculating machine. The resulting moving averages are listed in Table 20-2, column 4.

3. Divide each monthly item of original data by the corresponding 12-month moving average, and list the quotients as "Percent of Moving Average." In Table 20-2, column 2 divided by column 4 equals column 5. Division is preferable to subtraction here because seasonal variation tends to repeat itself from year to year with the same *relative* intensity. That is, a normal seasonal rise in a given month tends to remain at the same percent as the enterprise grows, even though the dollar value rise in this month increases with the size of the business. Since the 12-month moving average roughly describes the path of the trend and cyclical fluctuations combined, the percentages of the original data divided by this average represent primarily the seasonal-irregular components, as in the graphic method. That is, actual sales = trend (T) \times cycle (C) \times seasonal (S) \times irregular (I) components in our time series model. (Trend is expressed in the original unit, such as dollars, while the other components are stated as percents.) Then, in step 3, $TCSI/TC = SI$, and averaging the SI ratios in the same month for different years (step 4) cancels out most of the I factor.

4. Compute the modified mean of the percents of moving averages for a given month in the various years, omitting the highest and lowest values as being dominated by irregular factors, exactly as in the graphic method.

The percents in Table 20-2, column 5, are grouped in Table 20-3. The highest and lowest figures in each column are then crossed out, as before, and the remaining four values are totaled and divided by 4 to give the *modified means*, or preliminary seasonal indexes.

5. Since the 12 modified means total 1,203.4 rather than 1,200 (last column), each one is multiplied by $1,200/1,203.4$ to yield the final seasonal indexes shown in the row below. These indexes total 1,200 and therefore average 100 percent.

Since steps 4 and 5 are both the same as in the graphic method, Table 20-3 is quite similar to Table 20-1, and a graph of the figures in Table 20-3 (not shown here) would show nearly the same pattern of seasonal indexes and seasonal irregularities as in Chart 20-3. The seasonal indexes obtained by the two methods are compared at the bottom of Table 20-3. The average absolute difference between the two is only 0.2 points for the 12 months, which is trivial, since seasonal indexes are only accurate to within about one point, unless more refined methods are used.

6. In order to adjust the data for seasonal variation (to eliminate its

Table 20-3

PERCENTS OF 12-MONTH MOVING AVERAGES
AND COMPUTATION OF SEASONAL INDEXES

SEARS, ROEBUCK SALES, 1960-1965

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Totals
1960	78.2	71.0	83.1	104.1	97.6	103.5	95.9	104.1	99.9	108.4	100.7	149.4	
1961	75.2	68.8	90.6	90.8	100.7	104.2	96.4	104.6	103.6	105.4	111.8	144.7	
1962	77.9	69.0	88.7	96.5	103.5	102.3	94.6	104.0	98.8	104.3	113.0	145.4	
1963	79.1	69.5	89.9	97.8	102.4	102.1	96.8	106.2	97.9	102.5	110.7	150.5	
1964	78.7	73.5	89.7	95.5	100.2	103.1	97.0	102.6	98.2	106.7	111.2	152.0	
1965	77.6	71.7	87.1	98.1	100.3	97.2	98.2	102.0	102.1	103.5	114.9	152.3	
Total, middle four	309.4	281.2	355.4	387.9	403.6	411.0	386.1	415.3	399.0	419.9	446.7	597.6	
Modified mean	77.4	70.3	88.8	97.0	100.9	102.8	96.5	103.8	99.8	105.0	111.7	149.4	1,203.4
Seasonal index	<u>77.1</u>	70.1	88.6	96.7	100.6	102.5	96.3	103.5	99.5	104.7	111.4	149.0	1,200.0
Seasonal index (graphic)*	77.3	70.0	88.3	96.6	100.1	102.6	96.6	103.6	99.6	104.9	111.5	148.9	1,200.0
Difference	-0.2	0.1	0.3	0.1	0.5	-0.1	-0.3	-0.1	-0.1	-0.2	-0.1	0.1	

* From Table 20-1.

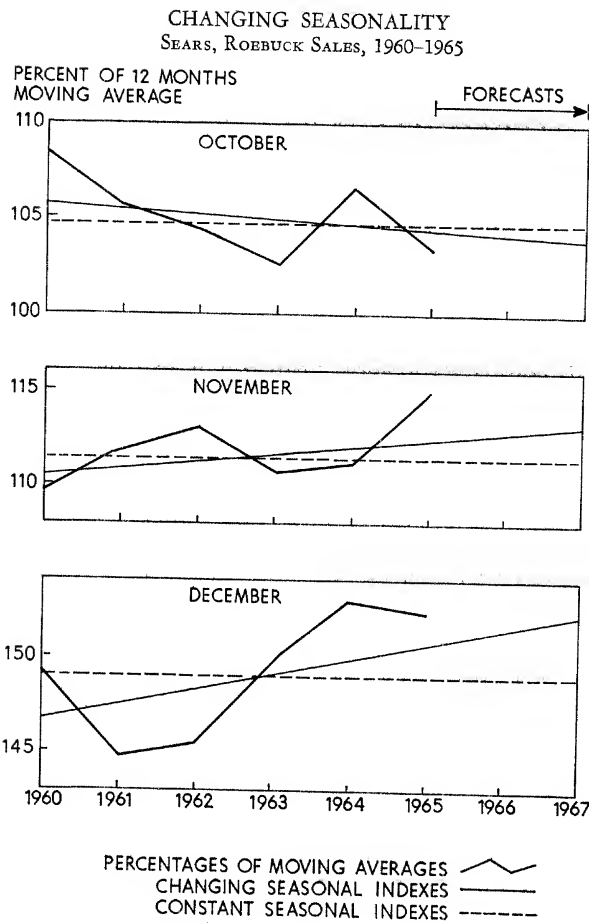
effects), divide the actual sales by the seasonal indexes. Thus, in January 1960, actual sales of \$271.3 million (Table 20-2) divided by 77.1 percent (Table 20-3) give \$351.9 million as the sales *adjusted for seasonal variation*. That is, $TCSI/S = TCI$. These figures are not listed here, since their graph would be almost identical with the dashed line in Chart 20-2 showing sales adjusted by the graphic method.

Changing Seasonality

Seasonal rhythm may change gradually over a period of years. Thus, Sears, Roebuck may be able to boost its Christmas sales, relative to other seasons, from year to year. New customs, such as increasing vacation travel in summer, stimulate many activities in this season. This gradual change in seasonal behavior is called *changing* (moving or progressive) seasonality, as opposed to the "constant" seasonality discussed above.

Changing seasonality may be measured as follows in either the graphic or moving-average method: (1) set up 12 small charts with the vertical scale marked "Percent of Trend-Cycle" or "Percent of 12-Month Moving Average," and mark the years on the horizontal scale. Either arithmetic or ratio charts may be used. Plot the January percents from Table 20-1 or Table 20-3 in the first chart as a time series, the February percents in the second chart, and so on. Then, if the January points show a sustained upward or downward drift over the years, draw a smooth, freehand trend curve through the plotted points. Now, read off the preliminary seasonal indexes from the trend curve, a

Chart 20-4



Source: Table 20-3

different index for January in each year. Correct the 12 indexes in each calendar year to average 100 percent if necessary, as in step 5 above.

Chart 20-4 shows the percents of 12-month moving averages for October, November, and December, from Table 20-3, plotted as time series. The October percents appear to drift downward, while those for November and December follow a rising trend. Therefore, we have drawn sloping freehand curves through these panels to smooth out the irregularities and thus determine the preliminary changing seasonal indexes over the years. The index is read from this curve each year, rather than using the constant seasonal index from Table 20-3, which is drawn as a horizontal line. The curves have been projected ahead to 1967 for use in forward planning.

This trend fit is justifiable provided there is some known explanation for the shift and a long enough period of years is included to be sure that our slope does not represent merely a random run. In this case it appears that the Christmas season is expanding relative to certain other months, such as October, although the evidence of only six years is not conclusive.

To check this tendency over a longer period, Table 20-4 presents constant seasonal indexes for three six-year periods since World War II, all computed by the moving-average method. It appears that the periods January–April and September–November have declined in importance, while the formerly depressed summer months and December have expanded. This confirms the short-term trends in Chart 20-4 for October and December, but not for November. For more detailed analysis,

Table 20-4

THE CHANGING SEASONAL PATTERN OF SEARS, ROEBUCK SALES
(Constant Seasonal Indexes in Three Periods, 1946–1965)

Period	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1946–1951	81.8	71.9	93.5	98.7	98.7	98.9	87.1	97.5	105.7	109.9	114.9	141.4
1953–1958	77.0	70.2	86.4	96.8	104.8	105.8	94.4	102.3	101.1	107.0	109.6	144.8
1960–1965	77.1	70.1	88.6	96.7	100.6	102.5	96.3	103.5	99.5	104.7	111.4	149.0

we should extend Chart 20-4 to cover all months and a longer period of years.

Changing seasonal measurement is recommended for refined analysis, since it takes into account gradual changes in seasonal behavior. However, it still does not allow fully for cyclical changes in seasonality, such as the pickup in the slack season during cyclical booms, or abrupt changes, such as those caused by war. Disruptions can best be avoided by simply omitting the abnormal periods in computing the seasonal indexes. Furthermore, changing seasonal indexes are cumbersome because they differ for each month of each year. For ordinary purposes, therefore, the use of constant seasonal indexes for homogeneous periods of years should be adequate.

Use of Electronic Computers

Electronic computer programs for measuring seasonal variation have been developed in recent years to speed the computations and permit various refinements of technique. Two of the principal methods are the

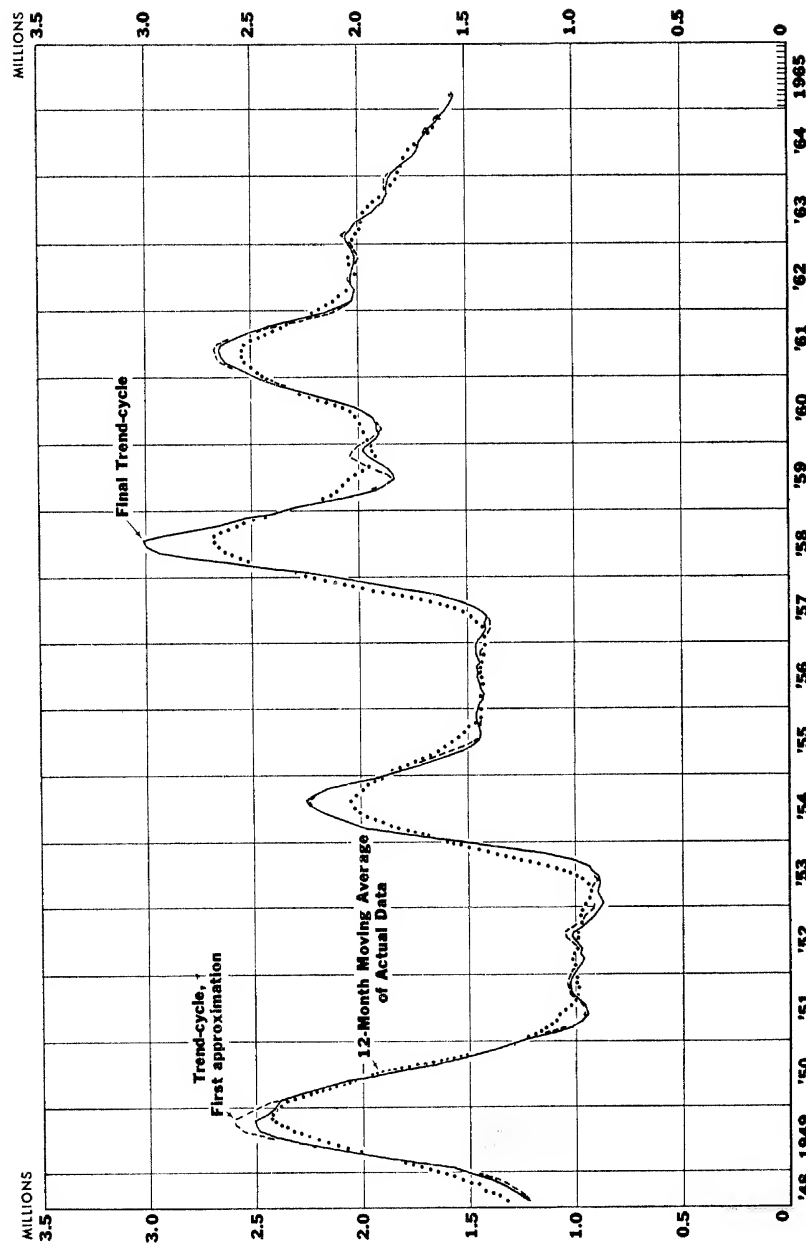
Census II Seasonal Adjustment Program¹⁰ and the BLS Seasonal Factor Method.¹¹ The Census II program is available in FORTRAN IV language, which can be used on many medium- and large-scale electronic computers. The BLS program is adapted to the IBM 1401 and 1460 tape systems. The typical run will require less than five minutes of computer time. Both methods are based on the ratio-to-12-month-moving-average method, using changing seasonal indexes, but the programs offer a variety of optional refinements, summary measures, and tests of significance.

The Census II method has these important features: (1) The preliminary calendar correction can be performed by correlating the original series with the number of times each day of the week occurs in each month, rather than by having to introduce explicitly the number of working days in the month. (2) The series is then adjusted for seasonal variation by the ratio-to-centered-12-month-moving-average method. (3) The adjusted series (TCI) is then smoothed by a weighted moving average of 9, 13, or 23 terms (depending on how irregular the series is), in order to smooth out irregularities and provide a revised trend-cycle curve. Chart 20-5 (taken from the BLS method) shows that this type of trend-cycle curve is much more sensitive to cyclical movements than the original 12-month moving average, as applied to unemployment data for 1948-1965. In particular, the unemployment peaks of 1949, 1954, 1958, and 1961 are much more pronounced than those shown by the 12-month moving average. (4) The original daily averages are then divided by this new trend-cycle base and the seasonal measurement process is repeated as before. (5) The seasonal-irregular ratios for a given month in different years are smoothed by a weighted moving average (obtained by taking a three-term average of a five-term moving average) to estimate the changing seasonal indexes. (6) Extreme values are given reduced weight or no weight, depending on how many standard deviations they depart from the norm. (7) A set of summary measures is prepared, such as the percent contributions of the trend-cycle, calendar, seasonal, and irregular factors in a time series, and the ratio of the average irregular component in month-to-month changes to the average trend-cycle component. Various tests of signifi-

¹⁰ See U.S. Bureau of the Census, "The X-11 Variant of the Census II Seasonal Adjustment Program," Technical Paper No. 15 (November 1965), summarized in *Business Cycle Developments* (October 1965), pp. 57-71. These sources include a sample printout and bibliography.

¹¹ U.S. Bureau of Labor Statistics, May 1966.

Chart 20-5
DEVELOPMENT OF TREND-CYCLE COMPONENT
UNEMPLOYED MEN* IN THE UNITED STATES, JULY 1948-JUNE 1965



* Age 20 and over.
SOURCE: U.S. Bureau of Labor Statistics, *The BLS Seasonal Factor Method* (1966), p. 6.

cance are also provided. (8) The results are printed out graphically on a chart.

Therefore, the electronic computer carries the ratio-to-moving-average method through more refinements than would otherwise be feasible. Furthermore, seasonality can be analyzed in far more economic time series than was formerly possible.

Electronic computers cannot handle certain problems such as abrupt changes in plantwide vacation schedules or the shifting dates for offering new automobile models. These situations should be adjusted by hand before the data are put into the computer, or else the series should be broken at the point of discontinuity and the two segments analyzed separately. Computers provide speed and precision of results in the hands of the skilled analyst, but they still do not take the place of human judgment.

Which Method to Use?

The following suggestions may be helpful in selecting an appropriate method for measuring seasonal variation:

1. The graphic method is recommended as a short cut, since it substitutes graphic measurements for the three laborious steps (2, 3, and 6) of the moving-average method. The freehand trend-cycle curve can follow cyclical movements more closely than the 12-month moving average, if drawn with skill and judgment, particularly when revised to follow the seasonally adjusted data. The graph also affords a visual check on each step, revealing irregularities in the data and allowing necessary variations in technique.

2. The moving-average method has the advantage of being a standard, objective procedure that can be performed by clerical labor with a hand calculator and adding machine. It is the most commonly used of the many simple arithmetic methods proposed for analyzing seasonality. Like the graphic method, its results are usually accurate enough for ordinary purposes.

3. Electronic computer methods provide both the greatest time saving and the most accurate seasonal measurement, when many series are to be analyzed, and the program and computer are available. Such programs, however, are complex and require a sophisticated analyst to select the appropriate options and to interpret results.

Other Methods of Taking Seasonality into Account

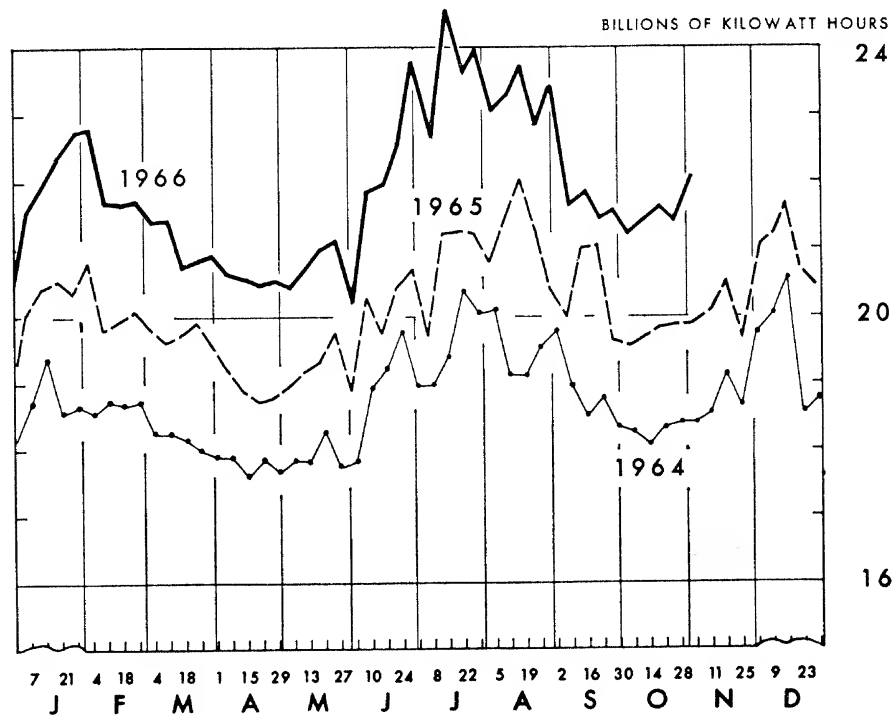
There are several commonly used methods of allowing for seasonality without actually measuring it:

1. Seasonal movements are sometimes referred to merely in directional terms. For example, "Retail sales made a seasonal gain in September over the August level." This statement, however, does not say whether the gain was more or less than the normal seasonal amount and how much it differed. It would be more meaningful to say: "Retail sales gained 8 percent in September over the August level, after allowance for the usual seasonal increase."

2. The common practice of comparing a month with the same month a year ago serves to eliminate the seasonal factor common to both months. This usage, however, may still distort the cyclical picture for either of two reasons: (a) The current month is judged in comparison with a single historic month that might have been erratic itself. Thus, the statement "Production in March was 3 percent above a year ago" appears favorable, but it might represent an unfavorable situation if March last year was unduly depressed. (b) The comparison with a

Chart 20-6

ELECTRIC POWER PRODUCTION



SOURCE: *Federal Reserve Chart Book* (November 1966). This source also charts the seasonally adjusted data, which clarify nonseasonal movements.

year ago ignores the trends of the past 11 months. For example, Sears, Roebuck sales in January 1961 were above those of January 1960. This report appears favorable, but it would have been more significant to note that seasonally adjusted sales had been declining during the recession of late 1960, as shown in Chart 20-2. Similarly, sales in each of the six months of April-September 1966 reached a new all-time record for this month, but they fell below expectations, based on our forecast of normal seasonal behavior and a continued trend-cycle rise.

3. Plotting weekly or monthly data for several years above each other on a tier chart with the horizontal scale extending from January to December enables one to compare current tendencies with those in the same seasons of other years without any calculations. But the comparison with several such years is apt to be confusing and offers no precise adjustment for the seasonal factor. In Chart 20-6, for example, the general level of 1966 electric power production is obviously above that of the two previous years, but the weekly nonseasonal comparisons are not clear. In particular, was the decline in production during August and September 1966 more or less than the usual seasonal amount?

These methods are sometimes useful for simple presentation. For careful analysis, however, seasonal indexes should be computed as described earlier in the chapter.

USE OF SEASONAL INDEXES IN SHORT-TERM FORECASTING

Seasonal indexes play an important part in short-term business planning. Chart 20-2 shows how Sears, Roebuck sales can be forecast (at the end of 1965) for each month of 1966 by projecting the trend-cycle curve and multiplying these values by the seasonal indexes. The same technique, of course, can be applied to individual products or departments, as well as to total sales.

In order to project the trend-cycle component of sales, let us break it down into three elements: (1) the secular trend in deflated sales, (2) price changes, and (3) cyclical movements. Now the growth in sales has been about 12 percent in each of the years 1963, 1964, and 1965. About 5 percent of this represents the trend in deflated sales (see page 492), 1 percent represents price inflation (Table 19-1), and the remaining 6 percent reflects primarily the general business cycle expansion. Since the company distributes durable goods on a nationwide scale, its sales are especially sensitive to changes in U.S. personal income and credit rates.

At the close of 1965, the forces of secular growth and price inflation promised to continue unabated in 1966, so we have projected these

elements at their combined historic growth rate of 6 percent. The cyclical outlook, however, indicated some slowdown in the 1962–65 rate of expansion. In particular, *Predicasts* forecast a reduction in the growth rate of disposable personal income in the latter half of 1966; United Business Service predicted a 6 percent rise in total retail sales in 1966, compared with 8 percent the previous year; and *Business Week* estimated only a 4.5 percent increase in retail sales for 1966. Furthermore, the long boom had built up consumer stocks of durable goods, and tight money had made credit purchases more difficult.

Therefore, we have made a judgment estimate that the recent cyclical growth rate of 6 percent a year would taper off to 3 percent by the end of 1966. This is rough; a complete cyclical forecast would require detailed economic analysis beyond the scope of this book. (We do treat indicators of cyclical turning points in Chapter 21, and regression analysis for predicting sales from their relation with personal income, number of stores, and other predictable factors in Chapters 22–24.)

Combining the effects of secular growth, price inflation, and cyclical expansion, therefore, we estimated that Sears, Roebuck annual growth rate would decline from 12 percent at the beginning of 1966 to 9 percent at the end of the year.

In Chart 20–2 we begin the Sears, Roebuck trend-cycle projection at the seasonally adjusted average of the last quarter of 1965 (i.e., \$603 million, plotted at the middle month of November) since the 3-month period serves to iron out the irregularities of individual months. We then have extended the trend-cycle curve through 1966, beginning with the same 12 percent annual growth rate as in the past three years, but arbitrarily tapering off to a point in December 1966 only 9 percent above December 1965.

We can then read off from the chart the trend-cycle value for each month of 1966 and multiply this by the seasonal index to obtain a forecast. That is, $TC \times S = TCS$. (The irregular element cannot be estimated.) Alternatively, we have here used the graphic method—placing the 100 percent mark of the measuring strip on the trend-cycle curve of the chart and marking the forecast sales opposite the seasonal index on the strip to repeat the seasonal pattern shown in prior years.

Our 1966 forecast is shown in Chart 20–2 as a dotted line, together with actual sales for January–September 1966 (solid line) plotted later as a check on this projection. For the first three months of 1966 the forecast proved quite accurate, but thereafter actual sales fell below the projections, apparently because of the extremely tight money situation

and the sharp stock market break that occurred at that time. This example illustrates the need for management to review such a forecast at least quarterly and to revise it as required by new developments.

The error of the forecast includes that of the trend-cycle projection (which increases with time) and that of the irregularity in the seasonal element itself, which can be estimated from the scatter of the arrays in Chart 20-3. When seasonal fluctuations are large and regular, and short-term cyclical movements are mild, as in retail trade generally, short-term forecasting is relatively accurate.

SUMMARY

Seasonal variations are regular rhythmic movements within a period of one year resulting from the weather and from man-made conventions such as holidays. They affect nearly all economic processes in varying degrees, particularly at the point of origin and the point of consumption. Seasonal variations may change in character over the years. However, seasonal fluctuations are much more regular than cycles, and so they can be measured and projected more accurately. Regular rhythms also occur within a quarterly, monthly, weekly, or daily period. Finally, the calendar itself causes quasi-seasonal variations in monthly and weekly data, since the number of operating days varies from one month or week to the next.

Adjustment for calendar variation is made as a preliminary step in seasonal measurement in order to eliminate fluctuations in the data caused by the varying length of the working month. The data are divided by the number of operating days in each month to place the series on a uniform daily average basis. The number of operating days must be determined separately for each industry and area. Weekly data are adjusted only for holidays, the number of weekdays being constant.

Seasonal variation is measured for the purpose of understanding past fluctuations, forecasting and budgeting, or adjusting data in order to reveal cycles. The seasonal pattern is best described by seasonal indexes that represent the average value for each month related to the average of all 12 months as 100 percent. The period analyzed should be long enough to average out peculiarities in individual years, but abnormal periods should be omitted.

Several methods of computing seasonal indexes are described. The graphic and moving-average methods are summarized in the table, with symbols to indicate how the trend (*T*), cycle (*C*), and irregular (*I*) factors are eliminated to isolate the seasonal index (*S*).

Step	Graphic Method	Moving-Average Method	Shows
1	Plot on ratio chart	Plot on ratio chart	<i>TCSI</i>
2	Draw freehand <i>TC</i> curve	Compute 12-month moving average	<i>TC</i>
3	Read ratios of data to <i>TC</i> from measuring strip	Divide data by moving average	<i>SI</i>
4	Compute modified means of ratios for each month	Compute modified means of ratios for each month	<i>S</i> (preliminary)
5	Multiply indexes by 1,200 over their sum	Multiply indexes by 1,200 over their sum	<i>S</i>
6	To adjust for seasonality, shift plotted data the distance from seasonal index to base line of measuring strip	To adjust for seasonality, divide data by seasonal indexes	<i>TCI</i>

Results can be improved by redrawing the trend-cycle curve through the seasonally adjusted data and repeating the seasonal measurement process.

If the seasonal pattern changes over the years, changing or moving seasonal indexes can be computed in either method by plotting the ratios for each month in step 3 chronologically and reading the preliminary indexes from freehand trend curves drawn through these plots.

Electronic computer programs such as Census II and BLS greatly speed the necessary calculations and permit several refinements in technique, such as calendar adjustment from internal evidence, improved trend-cycle estimates using weighted moving averages, reduced weights for extreme items, computation of changing seasonal indexes, and various summary measures and tests of significance.

The methods compare as follows: the graphic method is quick, flexible, and affords a continuous check on operations, while the moving-average method is objective and can be performed by clerical labor on hand calculators. Electronic computer programs are recommended where many series are to be treated, since they give fast and accurate results in the hands of a skilled analyst.

Seasonality is sometimes taken into account without actual measurement by means of (1) qualitative description, (2) comparing a month with the same month a year ago, or (3) plotting several years on a tier chart with the same monthly time scale. These devices are useful for simple presentation, but seasonal indexes are needed for refined analysis.

To make a short-term forecast, project the trend-cycle curve (see

cyclical forecasting) and multiply these values by the seasonal indexes each month (i.e., $TC \times S = TCS$) or lay off these indexes from the TC curve with the measuring strip graphically.

PROBLEMS

1.
 - a) Select and plot a series of monthly data dominated by seasonal movements. The graph may be traced on a blank sheet placed over a chart in a current publication. Do not use textbook examples.
 - b) Describe the seasonal characteristics: Is the seasonal amplitude wide or narrow? Is the seasonal pattern regular or irregular? What are the high and low months and the seasonal tendency of other months? Give reasons for these movements.
2. Which of the following should be changed to an average daily basis, and which should not? Explain in each case.
 - a) Monthly data on average sales per sales person in a chain of women's apparel stores.
 - b) A monthly record of the stocks of a department store.
 - c) The total loans of a commercial bank on the last day of each month.
3.
 - a) List, from Moody's or Standard and Poor's reports, Sears, Roebuck sales for the first five months of this year or last year.
 - b) Adjust these sales to a daily average basis, counting Saturday as $1\frac{1}{2}$ days and omitting Sundays, January 1, and May 30. (See calendar.)
 - c) Plot the actual sales and daily average sales on a small chart, using two scales.
 - d) How does the calendar adjustment affect month-to-month movements?
4.
 - a) Define "seasonal index." Distinguish between constant and changing seasonal indexes.
 - b) Having computed seasonal indexes, describe briefly how to make a seasonal forecast.
 - c) A chart is captioned "Adjusted for Seasonal Variation." Explain.
 - d) Why is it sometimes necessary to adjust monthly data for calendar variation before measuring seasonality?
5. Seasonal indexes of sales for the Ace Products Company are January, 97; February, 89; March, 101; April, 104; May, 120; etc.
 - a) Company sales increased from \$2,910,000 in January 1967 to \$2,964,000 in April of the same year. What was the percent change in the seasonally adjusted sales between January and April?
 - b) The company treasurer has forecast sales of \$36 million for the next calendar year. He believes that by May the trend-cycle component should be about 5 percent above the average monthly level. Based

upon his assumptions, what is the treasurer's sales forecast for the month of May?

Problems 6 to 8 utilize the following data:

COASTAL CEMENT COMPANY					
PRODUCTION OF PORTLAND CEMENT, 1963-1967, Thousands of Barrels					
YEAR	QUARTER				ANNUAL AVERAGE
	First	Second	Third	Fourth	
1963	100.3	148.5	147.6	128.7	131.3
1964	111.5	162.9	164.6	147.2	146.6
1965	142.5	171.2	170.8	162.5	161.8
1966	151.0	174.8	167.6	155.1	162.1
1967	147.3	168.8	167.7	153.6	159.4
Total	652.6	826.2	818.3	747.1	761.2
Quarterly average	130.5	165.2	163.7	149.4	152.2

6. *a)* Compute indexes of seasonal variation for the cement production data above by the graphic method.
 - b)* Adjust this series graphically for seasonal variation.
 - c)* Forecast cement production graphically for the four quarters of 1968, extending your trend-cycle curve freehand.
7. *a)* Compute indexes of seasonal variation for the cement production data above by the moving-average method, centering the moving average on the third quarter. Use these additional production figures: 1962, third quarter, 156.0 thousand barrels; fourth quarter, 132.2; and 1968, first quarter, 137.3 thousand barrels.
 - b)* How much do these indexes differ from those of the graphic method? Give reasons for the differences.
 - c)* Adjust this series arithmetically for seasonal variation and plot the results. What is the purpose of this adjustment?
 - d)* Forecast cement production in the second quarter of 1968, assuming a trend-cycle decline of 2 percent from the first quarter.
8. *a)* What factors determine whether constant or changing seasonal indexes should be computed?
 - b)* How does the computation of a changing seasonal index differ from that of a constant seasonal index?
 - c)* Is there evidence of changing seasonality in cement production (Problem 6 or 7 above)? Present small charts of each of the four quarters to support your answer.

9. As an analyst with the General Oil Company, you wish to measure the seasonal variation in the company's gasoline sales by the graphic method, using the following data:

GENERAL OIL COMPANY							
GASOLINE SALES, DAILY AVERAGES IN HUNDREDS OF BARRELS							
	1961	1962	1963	1964	1965	1966	1967
January.....	252	264	269	274	330	327	361
February.....	271	263	278	295	330	355	398
March.....	264	283	298	318	336	348	382
April.....	287	300	320	334	357	397	407
May.....	287	307	321	359	374	398	406
June.....	317	340	351	368	406	410	452
July.....	298	328	342	377	399	429	438
August.....	320	335	355	376	408	428	
September.....	304	342	344	367	380	416	
October.....	298	298	319	348	401	411	
November.....	275	311	320	332	349	376	
December.....	296	292	308	324	344	387	
Average.....	289	305	319	339	368	390	

- Plot the data on a one-cycle ratio chart; draw a trend-cycle curve through the 1961-1966 annual averages (extended through 1967), and determine the twelve seasonal indexes by means of a measuring strip.
 - Describe briefly the typical seasonal behavior in the company's sales. Is the seasonality regular or irregular?
 - Forecast gasoline demand for the next four months (August-November 1967) by laying off the seasonal indexes from your measuring strip above or below the extended trend-cycle curve on the chart. Plot your forecast as a dashed line, and the actual figures below (determined later) as a solid line to compare the results. Actual sales were: August, 433; September, 438; October, 411; November, 392.
 - What was the probable cause of the error in your forecasts over this four-month period?
 - Adjust the data for seasonal variation graphically and plot the results in red. Describe the principal nonseasonal movements in gasoline demand over this period. Which of these movements dominates the adjusted series—trend, cycles, or irregular fluctuations?
10. In order to analyze the factors affecting gasoline sales of the General Oil Company, you decide to compute indexes of seasonal variation for the data in Problem 9 by the moving-average method. You first compute a 12-month moving average for each month, and then divide the original sales by these averages; obtaining the following percentages:

GENERAL OIL COMPANY

MONTHLY GASOLINE SALES AS PERCENTAGES OF 12-MONTH MOVING AVERAGES

	1961	1962	1963	1964	1965	1966	1967
January.....	91.5	89.0	86.1	83.1	92.9	86.7	89.4
February.....	97.8	88.0	88.7	88.9	92.2	93.5	98.1
March.....	94.5	94.1	94.7	95.4	93.4	91.0	91.5
April.....	101.8	99.2	101.2	99.3	98.4	103.4	101.6
May.....	101.0	101.0	101.3	106.4	102.4	103.1	98.5
June.....	110.3	111.3	110.2	108.7	110.7	105.5	108.1
July.....	102.8	107.5	107.4	110.2	108.4	109.6	108.1
August.....	110.6	109.4	111.1	108.9	110.8	108.5	
September.....	104.8	111.1	107.1	105.5	102.6	104.6	
October.....	102.1	96.3	98.9	99.6	107.6	103.0	
November.....	93.8	100.2	98.4	94.6	93.1	93.9	
December.....	100.4	93.6	94.1	91.7	91.5	96.2	

- a) If the original data represent $T \times C \times S \times I$ (trend \times cycle \times seasonal \times irregular forces), what types of fluctuations do the data in the above table primarily represent? How were these elements derived from the original figures?
 - b) Compute a modified mean of these percents for each of the 12 months (omitting the highest and lowest percent in each case as being the most erratic) to average out the irregular elements. Then multiply these means by (1,200/their total), if necessary, so that they will average 100. List the resulting seasonal indexes rounded to the nearest whole number.
 - c) In July 1967 the company economist predicts that a cyclical recession during the balance of the year will offset the usual secular trend growth. On this assumption, forecast daily average gasoline sales for November 1967 based on the normal seasonal change from July (the latest month available). Give percent error of forecast, compared with actual figure of 392 thousand barrels daily average in November.
 - d) You wish to analyze the change in gasoline sales between February and July 1967. Actual sales increased from 398 to 438, or 10 percent, in this period. Adjust the data in these two months for seasonal variation and compute the percentage change in the adjusted figures.
 - e) Show how the adjusted February and July figures were derived, in terms of the $TCSI$ concept, and explain the significance of the change in the adjusted demand.
11. Gasoline demand is said to be less seasonal than formerly, since people in colder areas who once stored their cars during the winter now drive the year round, and vacation trips that were formerly confined to the summer months are now made throughout the year. Do the figures in Problems 9 and 10 confirm this claim? That is, does gasoline demand in a winter month, expressed as a ratio to the average month tend to rise, and does the ratio for a summer month fall correspondingly over the years? Test this hypothesis of changing seasonality for the two months February and June as follows:
- a) Plot the February and June percentages-of-moving averages from Problem 10 on two panels of an arithmetic chart.

- b) Draw a freehand trend line through each of these diagrams, ignoring erratic points.
 - c) Do these charts support the claim that gasoline demand is becoming less seasonal? Explain.
 - d) Read off from your trend lines and list the changing seasonal indexes for February and June 1967.
12.
 - a) Cite the one chief advantage of graphic methods and of arithmetic methods, respectively, in seasonal analysis, and explain your choice.
 - b) In what type of study may the electronic calculator method be preferable?
 - c) How could you measure the *irregularity* of seasonal fluctuations in your business?
13.
 - a) Find a series of recent monthly data that is published both with and without seasonal adjustment in *Survey of Current Business*, *Federal Reserve Bulletin*, or some other source. Discuss the latest monthly figure in terms of (1) the percent change in the unadjusted value over a year ago and (2) the relation of the seasonally adjusted value to those of recent months. Compare these two methods of taking seasonality into account.
 - b) Find a weekly business indicator presented as a tier chart for the past several years and describe its recent behavior, indicating what component types of fluctuations can be distinguished. (One source is the *Federal Reserve Chart Book*.)

SELECTED READINGS

Readings for this Chapter are included in the list which appears on page 549.

21. CYCLICAL AND IRREGULAR FLUCTUATIONS

CYCLICAL FLUCTUATIONS, or alternations between expansion and recession, are of prime importance in short-term business analysis and planning.

Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own.¹

Business cycles have developed in modern industrialized countries having closely integrated business structures. The cycles are affected by factors outside business, such as wars, acts of government, and the size of crops, but it is the conditions within the business system itself that cause a protracted prosperity to give way to depression, and vice versa, in a roughly rhythmic fashion. Nearly all economic activities are affected by cyclical forces, but heavy industrial production and finance are most susceptible, and retail trade, personal service, and agricultural production are least affected.

The average length of business cycles in this country since 1919 has been about 4 years, of which the expansion phase has been over twice as long as the contraction phase. Table 21-1 shows the turning points of

¹This definition of Wesley C. Mitchell is used as the point of departure in the National Bureau of Economic Research studies in business cycles. See Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946), p. 3.

the general business cycle, averaged from thousands of individual series by the National Bureau of Economic Research.

In addition to the "short" cycle described above, some observers assert the existence of longer cycles, such as a 9-year intercrisis cycle and an 18-year residential building cycle. A conjunction of declines in these several cycles is said to cause major depressions. In any case, successive

Table 21-1
TURNING POINTS OF BUSINESS CYCLES
IN THE UNITED STATES, 1919-1961

TROUGH	PEAK	NUMBER OF MONTHS		
		Expansion	Contraction	Total
March 1919.....	January 1920	10	18	28
July 1921.....	May 1923	22	14	36
July 1924.....	October 1926	27	13	40
November 1927.....	August 1929	21	43	64
March 1933.....	May 1937	50	13	63
June 1938.....	February 1945	80	8	88
October 1945.....	November 1948	37	11	48
October 1949.....	July 1953	45	13	58
August 1954.....	July 1957	35	9	44
April 1958.....	May 1960	25	9	34
February 1961				
Mean duration.....		35	15	50
Median duration.....		31	13	44

SOURCE: National Bureau of Economic Research, reported in *Business Cycle Developments*, Appendix A, February 1967. This source also gives earlier turning points, beginning in 1854.

cycles vary so widely in amplitude (percent rise and fall) and pattern, as well as in length, that their prediction is extremely difficult.

Cycles in individual series also differ markedly in these respects from the general business cycle. Consider the cyclical swings of gross national product, aluminum, and coal production in Chart 19-2, as the major deviations from the trend lines. Gross national product is relatively insensitive to the cycle, since it contains many stable types of expenditures, such as interest payments, while aluminum production is extremely volatile, and coal is both moderate in amplitude and more sensitive to general business conditions than is aluminum. All three series, however, reflect the booms of the two world wars and the depressions of 1921 and 1932. The study of cycles is more crucial in "cyclical" or sensitive industries than in stable activities.

Irregular fluctuations in economic time series are caused by such

forces as unusual weather, labor strife, war, government intervention, and all forms of unpredictable events. These forces are of two types. The first group serves as "originating forces" in inducing or altering business cycle movements. War and its aftermath, for example, tend to produce the familiar boom and bust phases of a major peacetime cycle. A government public-works program may stimulate a similar cycle on a smaller scale. A protracted steel strike, on the other hand, creates a condition similar to cyclical depression in that industry. These forces are generally unpredictable, although many Washington "services" advise business on what the government is likely to do, and whether there will be war, strikes, large or small crops, etc., with partial success.

The second group of irregular factors comprises the host of miscellaneous forces that act in a more or less random fashion to give a plotted curve its familiar zigzag contour. These factors are usually numerous, unidentifiable, and unpredictable. The random element varies widely in different series, from nothing in the Federal Reserve rediscount rate to a major influence in the value of building permits issued.

The irregular component in a time series represents the residue of fluctuations after secular trend, cyclical, and seasonal movements have been accounted for. In practice, however, the cycle itself is so erratic and is so interwoven with irregular movements that it is impossible to separate them, except in smoothing out some of the random factors of the second type.

REASONS FOR MEASURING CYCLES

Three important purposes are served by isolating the cyclical, or cyclical-irregular, component in a time series.

1. Measures of past cyclical behavior are valuable aids in studying the characteristic fluctuations of a business. These measures will answer such questions as: How sensitive is this business to general cyclical influences? What is the typical timing, amplitude, and general cyclical pattern of the company's production, sales, inventories, or raw material prices? How do these factors compare with those of other companies or with the industry as a whole? Are there leads or lags compared with other series that would aid in forecasting?

The study of business cycles is also one of the major branches of economics. Today economists generally recognize the need not only of theory but also of accurate statistical measures in order to gain a clear understanding of this phenomenon. Hence, the National Bureau of Economic Research and other agencies have devoted years of study to this measurement.

2. Successful businessmen plan ahead; planning requires forecasting; and forecasting involves a knowledge of both typical and recent cyclical behavior. Measures of *typical* cycles are used in the "economic rhythm" school of forecasting, which projects past cycles ahead in periodic fashion. Such measures also appear in the "specific historical analogy" method of relating present conditions to those in a comparable period of the past and anticipating similar developments. Measures of *recent* cyclical behavior are necessary as a starting point in any kind of forecast. Articles may be found in almost any business journal, particularly around the first of the year, containing forecasts based on cyclical indicators.

3. Cyclical measures are useful tools in formulating policy aimed at stabilizing the level of business activity. Major efforts are being made by the federal government and by business to iron out the business cycle, since depressions are disastrous for the economy. The President's Council of Economic Advisers and the congressional Joint Economic Committee are important agencies that evaluate cyclical indicators as aids in devising safeguards against depression. Accurate cyclical measures are as necessary in planning preventive action as in anticipating what will happen without such action.

Despite the importance of business cycles, they are the most difficult type of economic fluctuation to measure. This is because successive cycles vary so widely in timing, amplitude, and pattern, and because the cyclical rhythm is inextricably mixed with irregular factors.

HOW TO MEASURE CYCLES

The standard method of isolating cycles in economic data is to eliminate seasonal, secular, and irregular movements as far as possible and to plot the residuals to show the cyclical fluctuations.² Not all of these movements, however, need to be eliminated in practice. The more pronounced a noncyclical factor, the more it tends to obliterate the cyclical pattern and the greater the need for its elimination. Thus, a wide seasonal swing, a steep trend, or a violently zigzag irregular contour requires adjustment more than if each of these factors were neutral. Ordinarily, the seasonal adjustment is the most important of the three. Frequently, *only* this adjustment is made in the data, together with some smoothing of random-type irregularities. This is because the

² A method of *averaging* the cycles in seasonally adjusted data is described in Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946), chap. 2; also Wesley C. Mitchell, *What Happens during Business Cycles: A Progress Report* (New York: National Bureau of Economic Research, 1951).

secular trend does not ordinarily obscure short-term cycles, and the adjustment for trend introduces an error arising from the fitting of the trend curve itself. Furthermore, cycles cannot be separated successfully from the sustained irregular movements caused by originating forces.

Annual data need be adjusted only for secular trend, since seasonal and short-term irregular fluctuations tend to cancel out in the yearly totals. Chart 19-6 shows the yearly deflated sales of Sears, Roebuck from 1926 to 1965, adjusted for trend. The cycles in the annual data were described on page 479. However, since cycles are of short-term duration, monthly data are usually needed to give a more detailed picture.

Graphic Adjustment

Cycles may be isolated graphically as follows:

1. Adjust the data for seasonal variation as described above. To illustrate, Chart 21-1 is reproduced from Chart 20-2 to show Sears, Roebuck sales adjusted for seasonality by the graphic method (dashed line).

2. Draw a freehand curve through the adjusted data, if necessary, to smooth out the zigzag irregularities and bring out the trend-cycle component in clear relief. The deviations above the curve should equal those below. This trend-cycle curve itself usually suffices for cycle analysis. Thus, the trend-cycle curve of Sears, Roebuck sales reached a peak in the first quarter of 1966 and then gave a warning signal by turning downward, whereas the unadjusted sales in Chart 20-2 might have misled management, since they rose sharply from February through August 1966 because of seasonal influences. (This curve can also be used in place of the preliminary freehand trend-cycle curve or 12-month moving average in recomputing the seasonal indexes, as described in Chapter 20 under "Revision for Greater Accuracy.")

3. The trend-cycle curve in Chart 21-1 can be adjusted further for trend by fitting a smooth trend curve (e.g., a logarithmic straight line) and laying off the vertical deviations of the trend-cycle curve from the trend around a horizontal line. The result is the cyclical component expressed as a percent of trend. This procedure is not shown here, since it was illustrated for Sears, Roebuck annual sales in Chapter 19, and the trend adjustment is not usually necessary for short-term analysis.

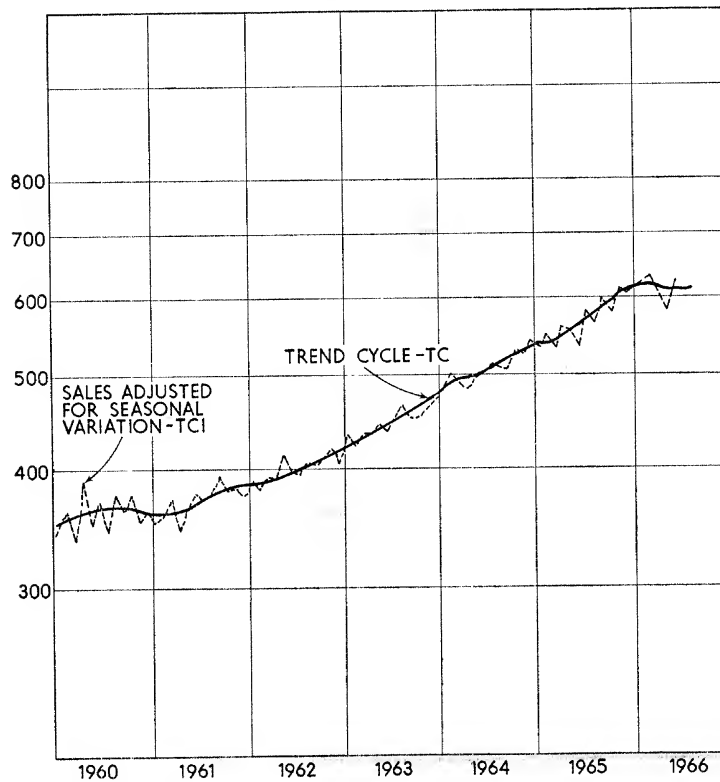
Arithmetic Adjustment

Cycles can also be isolated arithmetically in three steps:

1. Adjust the data for calendar and seasonal variation as described in the ratio-to-12-month-moving-average method.

Chart 21-1

TREND-CYCLE MOVEMENTS IN SEARS, ROEBUCK SALES, 1960-1966
 GRAPHIC METHOD
 Ratio Chart



SOURCE: Chart 20-2.

2. Compute a three-month moving average, if necessary, to smooth out short-term irregular movements. That is, the January-March average is plotted in the middle month, February; the February-April average is used for March; and so on. If the data are extremely erratic, a five-month moving average may be preferable. This results in a smoother curve but one which is less sensitive to month-to-month movements than the three-month moving average. Of course, irregular movements do not exactly offset each other every three or five months, so some of the irregularities remain in the smoothed curve. Ordinarily, the resulting trend-cycle values can be used for cycle analysis without further adjustment.

3. If it is desired to adjust for trend, fit an appropriate trend curve to the monthly data by least squares and divide the seasonally adjusted data by the trend values before computing the three- or five-month

moving averages. (However, the order of operations makes little or no difference.) That is, assuming that sales represent the product of $T \times C \times S \times I$,³ the seasonal adjustment is $TCSI/S = TCI$; dividing by the trend value gives $TCI/T = CI$; and a three- or five-month moving average cancels out part of the irregular movements to leave C as a residual. All steps can be performed by hand calculators.

We will not illustrate the arithmetic method of isolating cycles in Sears, Roebuck sales here, as we have already described step 1; step 2 is laborious; step 3 is usually unnecessary; and the TCI and TC curves resulting from steps 1 and 2, respectively, would be quite similar to those shown in Chart 21-1. The chief difference is that the short-term moving average would be somewhat more irregular, though more objective, than the freehand TC curve.

Computer Methods

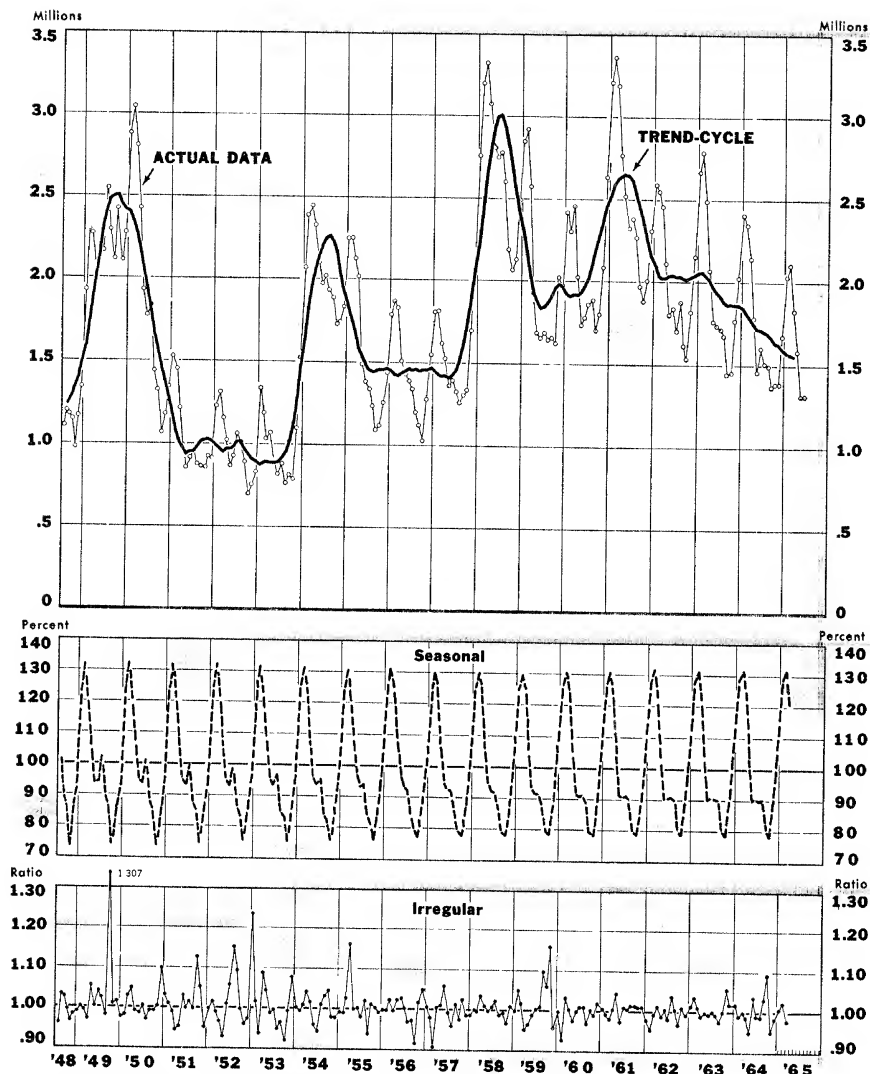
The electronic computer programs described in Chapter 20 not only adjust monthly or quarterly data for seasonality but also smooth out irregularities by means of a short-term moving average. An average of from one to six months is used in the Census II method, depending on the relative amplitude of the month-to-month *irregular* changes as compared with the *cyclical* changes in a series. That is, the number of "months for cyclical dominance" is computed as $MCD = \bar{I}/\bar{C}$, where \bar{I} is the average absolute irregular movement per month and \bar{C} is the average absolute cyclical change.⁴ This is the span in which the cumulative cyclical element in the series typically exceeds the irregular element. In a very irregular series such as liabilities of business failures, a six-month moving average is required for the cyclical element to dominate over the irregular movements. On the other hand, a single month's change in the Federal Reserve Board Index of Industrial Production typically contains a larger cyclical than irregular element, so the actual monthly figures are used without averaging several months.

Chart 21-2 illustrates the elimination of seasonality and the smoothing of irregularities in the number of unemployed men from 1948 to 1965, using the BLS computer method. The top panel shows the actual data and the final trend-cycle component, after eliminating the changing seasonal pattern and the irregularities depicted separately in the

³ This is $TCSI$, not $T + C + S + I$, since C , S , and even I tend to be more constant as percents than as absolute amounts. However, these factors can be added (or subtracted) on a ratio chart, since this operation is equivalent to adding the logarithms or multiplying the natural values.

⁴ \bar{C} includes the trend component, but this is negligible in one month. See *Business Cycle Developments* for a more detailed explanation.

Chart 21-2

TREND-CYCLE, SEASONAL AND IRREGULAR COMPONENTS
UNEMPLOYED MEN* IN THE UNITED STATES, APRIL 1948-JUNE 1965

* Age 20 and over.

SOURCE: U.S. Bureau of Labor Statistics, *The BLS Seasonal Factor Method* (1966), p. 2.

lower panels. Note how clearly the cycles of unemployment emerge in the trend-cycle curve, as compared with the actual data, which are dominated by strong seasonal-irregular influences. In particular, the peaks and troughs of the unemployment cycle occur at quite different times from those which appear in the actual data.

CYCLICAL FORECASTING

We can forecast monthly changes in a series for the next year by combining their trend, seasonal, and cyclical components. Projecting the trend and seasonal elements is a straightforward statistical process, but foretelling cyclical changes is much more difficult. Cycles are recurrent but not periodic; their expansion or contraction periods may be reversed at turning points that must be anticipated, or at least recognized in passing, for successful business planning. Also, unlike trends and seasonal movements, cycles in specific series are influenced by the general business cycle, so their prediction requires a study of the entire economy.

Naïve Methods

There are a number of "naïve" methods used explicitly or implicitly to foretell the near-term future. Some of these are as follows:

1. Assume that business next year will increase (or decrease) at the same percent rate as it did this year.
2. Assume that business next year will expand at the average secular trend rate of a number of past years.
3. Estimate that the duration of the current expansion or contraction phase of the cycle will equal the average of past cycles. However, individual cycles vary so widely in length of phase, as shown in Table 21-1, that the mean or median length of past cycles is of little predictive value.
4. Send a questionnaire requesting opinions on the business outlook to a broad mailing list of persons who may be interested, such as the subscribers to *Fortune* or the members of the Business and Economics Section of the American Statistical Association. Thus, from a quantity of casual replies one hopes to distill a precise forecast. The use of surveys to elicit a consensus of guesses is a widespread pastime in economic, political, and social affairs.

Some of these methods, particularly 1 and 2, prove more often right than wrong, since the usual estimate of continued rise reflects the long-term growth of the economy and the fact that cyclical expansions last longer than contractions.

Our cyclical forecast of Sears, Roebuck sales on page 524 was naïve in that it was a judgment estimate based on the consensus of views of professional economists on the outlook for personal income, retail sales,

and credit rates. However, the economists themselves undoubtedly used more sophisticated methods in arriving at their published forecasts.

Exponentially Weighted Moving Averages

A simple computer program can be used to forecast sales of a large number of products a few months ahead, for short-term planning and inventory control. The estimate is a moving average of past months, with weights declining exponentially. That is, the latest month is given the heaviest weight, and the weight for each preceding month is reduced by a constant percent. (The weights must total 1.) Such a procedure seems cumbersome, but it is actually simple for the computer, since all prior data can be summarized in a single number and only the latest month added to bring the moving average up to date. The result is often a reasonable estimate for the coming month since the moving average gives greatest weight to the latest month but still smooths out most irregularities by averaging a number of prior values. Trend and seasonal adjustments can also be incorporated in the program.⁵

The foregoing methods have the limitation of being based essentially on *past* trends rather than on future prospects. The most important function of business cycle forecasting, however, is not to predict a continuance of the current phase, but rather to recognize the turning points. The following methods may be useful for this purpose.

Lead and Lag Indicators

Most business processes move up and down roughly concurrently in the business cycle, but some are more sensitive than others, or represent earlier stages in production, and so reach their peaks and troughs before the aggregate indicators. Thus, the average work week of production workers in manufacturing responds more promptly to economic stimuli than does total nonagricultural employment. New orders for durable goods and construction contracts precede actual business expenditures for new plant and equipment. Common stock prices anticipate future changes in profits. Finally, sensitive commodity prices such as steel scrap move more promptly than composite nonfarm wholesale prices.

The Natural Bureau of Economic Research has selected a number of monthly and quarterly series that tend to lead the general business cycle

⁵ See Peter R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," in F. M. Bass *et al.*, *Mathematical Models and Methods of Marketing* (Homewood, Illinois: Richard D. Irwin, 1961), pp. 482-514. See also Robert G. Brown, *Smoothing, Forecasting, and Prediction of Discrete Time Series* (Englewood Cliffs, New Jersey: Prentice-Hall, 1963), chaps. 7 and 12.

at its turning points, another group that are roughly coincident in timing with general business, and some indicators that tend to lag.⁶ These are adjusted for seasonal variation and irregularities, as described on page 519, and are reported monthly in *Business Cycle Developments*, Chart 1. Thus, during a cyclical expansion, a marked downturn by a majority of the leading indexes gives a warning of a possible impending downturn in general business. If most of the coincident indexes then also turn down, this confirms the movements of the leaders, and if the lagging indicators follow suit, a general business recession is almost certainly in progress.

Unfortunately, none of these indicators is consistent in timing, and while most of them in fact have reversed direction at actual business peaks and troughs, they often give false signals because of minor intermediate movements, so they must be used with caution.

Diffusion Indexes

A diffusion index is also based on the principle that different processes in business reach their peaks and troughs at different times, but this device does not require identifying *which* particular series lead and which lag. A diffusion index is simply the percent of all seasonally adjusted series that are rising in a given month. (Sometimes a six- or nine-month span is also used.) Thus, if 60 out of 100 series increased in October over September, and 40 were stationary or declining, the diffusion index would be 60.

During the midexpansion period, perhaps 80 percent or more of all series are rising. But at the peak of aggregate activity, about half of the indicators of business volume will have turned down, while the other half are still rising, so that the diffusion index will cross the 50 percent line on the way down. Similarly, in midrecession the diffusion index may drop as low as 20 percent. But at the trough of general business, about half the series of business volume will have turned up while the other half are still declining, and the diffusion index will have risen to about 50 percent. Hence, a diffusion index signals a peak or trough in general business activity by crossing the 50 percent line on the way down or up. Theoretically, therefore, a diffusion index can lead the aggregates on which it is based by perhaps a quarter-cycle. Diffusion indexes are shown for many industries (e.g., new orders for durable goods in 36 industries) in *Business Cycle Developments*, Chart 2. Like the lead and lag indicators themselves, diffusion indexes usually mark

⁶ For a complete description, see National Bureau of Economic Research, *Business Cycle Indicators* (2 vols.; Princeton: Princeton University Press, 1961).

actual business cycle turning points very well, but often give false signals in crossing the 50 percent line because of short-term irregular movements.

Average Duration of Run

The diffusion indexes described above are unweighted in that each series counts the same. One weighting method is to assign each series in a given month a number from +6 to -6, depending on the number of months its trend-cycle component has moved up or down without interruption. Thus, if building contracts have moved up for six or more months through January it is marked +6, while if employment has declined two months since the last rise it is counted as -2. Then, these numbers are averaged for all series in a given month, and the resulting "average duration of run" series is plotted. It signalizes a peak or trough in business when it crosses the zero line, going downward or upward, respectively, just as the diffusion index does when it crosses the 50 percent line.

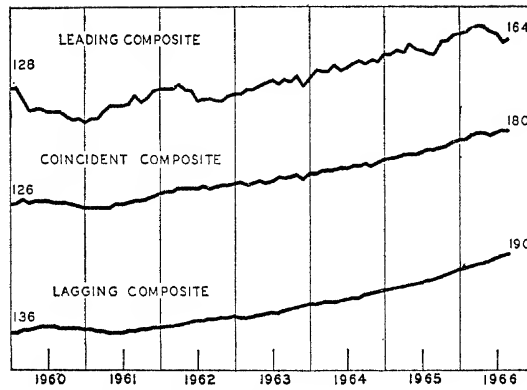
Chart 21-3 summarizes a group of leading, coincident, and lagging indicators, diffusion indexes ("Percentage Expanding") and average monthly duration, as compiled by Statistical Indicator Associates. As of October 1966, the leading indicators' composite had turned down, their percent expanding had dropped below 50, and their average monthly duration had sunk below zero. These signals warned of a possible cyclical peak to come in general business. However, none of the coincident or lagging indicators confirmed this downturn. Until they did, one should be cautious, but could not be assured of an imminent recession.

Surveys of Anticipations Data

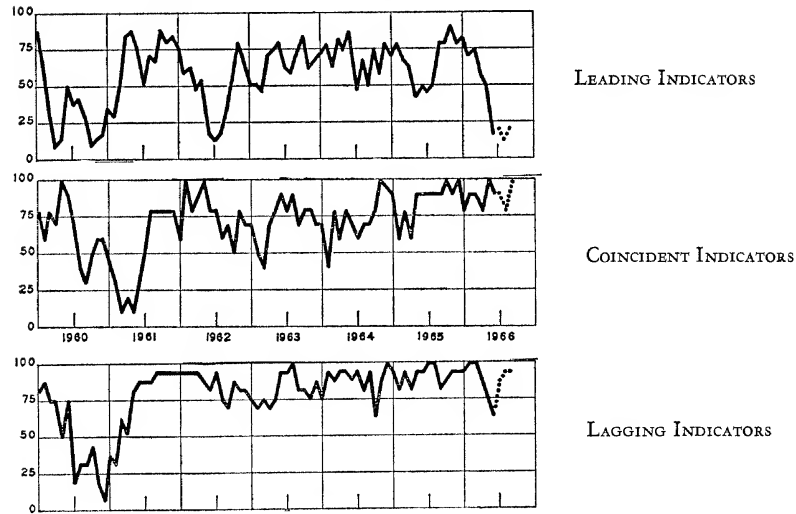
This method is based on the premise that businessmen, and to a less extent consumers, make forward plans for the expenditure of capital goods, and that a survey of these intentions will have forecasting significance. The surveys of businessmen's plans for new plant and equipment expenditures, conducted by the U.S. Department of Commerce-Securities and Exchange Commission and by McGraw-Hill, are widely followed. The National Industrial Conference Board surveys capital appropriations of large firms. The University of Michigan Survey Research Center and the U.S. Bureau of the Census canvass consumers' plans to purchase houses, cars, and durable equipment.⁷

⁷ See National Bureau of Economic Research, *The Quality and Economic Significance of Anticipations Data* (Princeton: Princeton University Press, 1960), for appraisal of these methods.

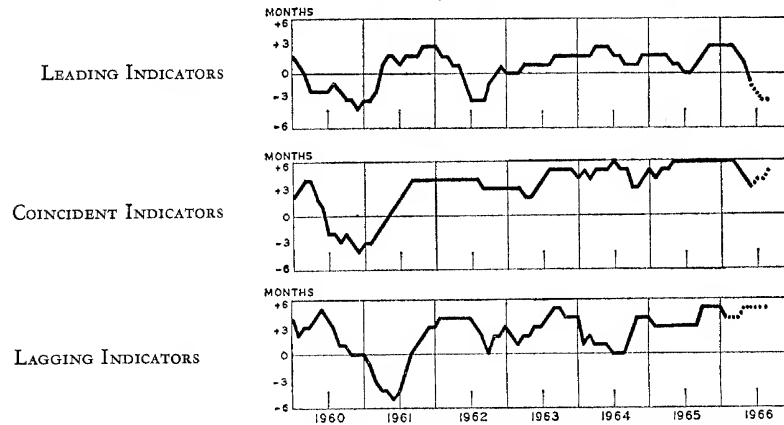
Chart 21-3
COMPOSITE INDEXES



PERCENTAGE EXPANDING



AVERAGE MONTHLY DURATION



SOURCE: Statistical Indicator Associates, North Egremont, Massachusetts.

Surveys of professional forecasters' opinions of course are valuable, as opposed to surveys of general mailing lists, which were classified under naïve methods above. Thus, the National Industrial Conference Board publishes the conclusions of an annual conference of leading forecasters. United Business Service summarizes the views of eight other financial services each month. The Federal Reserve Banks of Philadelphia and Richmond select and compile hundreds of forecasts early in the year. If you are confused by the multiplicity of expert opinions, just follow the consensus.

A RÉSUMÉ OF STATISTICAL METHODS IN FORECASTING

At this point we may summarize the statistical methods useful in business forecasting. Preliminary techniques of collecting data and presenting the results (Chapters 2 and 3) of course are essential. Sample survey methods (Chapter 14) are needed to survey the expectations of businessmen and consumers for the near-term future. Index numbers (Chapter 18) serve to summarize economic aggregates and their characteristics (e.g., diffusion indexes) as well as to make disparate series comparable. Time series analysis (Chapters 19–21) provide a means of projecting the secular trends, seasonal movements, and cycles of a business series to achieve a composite forecast. Finally, the correlation or regression analysis of time series (Chapters 22–24) will enable us to relate our own process (e.g., a company or industry sales) to some aggregate series (e.g., personal income) for which projections are available. Thus, *Predicasts* compiles forecasts for many economic aggregates and industry totals for up to fifteen years in the future, from many sources.

Not all the statistical methods used in short-term forecasting are needed in long-term forecasting. A long-term forecast, extending perhaps five or ten years in the future, typically involves a secular trend projection and regression analysis, to compare the series with basic economic aggregates. The long-term forecast is not concerned, however, with seasonal variation, nor is it possible to forecast the phase of the business cycle more than a year or two ahead. Surveys of anticipations or expectations, also, are generally not valid in the long run.

In short-term forecasting, which usually involves monthly estimates for the coming year, all the above statistical methods are applicable. In particular, it is useful to extrapolate the trend and seasonal movements of a monthly series by the methods described above, and then estimate by statistical and economic analysis whether the current phase of the business cycle is likely to continue or whether a turning point is in

prospect. Finally, the cyclical components of an individual series (e.g., industry sales) can be correlated with the cyclical elements in some basic series such as personal income, for which estimates are available. All the above methods can be carried out efficiently and comprehensively by electronic computers in large-scale analysis.

While statistical methods are necessary tools in business forecasting, they are not sufficient in themselves to complete the job. It is necessary to supplement the statistical results with a thorough economic analysis of cyclical and growth factors at the national, industry, and company levels. Accordingly, the corporate staff specialist responsible for forecasting is more often called a business economist than a statistician. The economics of forecasting of course lies beyond the scope of this book.⁸

SUMMARY

Cyclical fluctuations are the rhythmic movements of alternating prosperity and depression that have developed in industrialized economies. The average length of the short cycle is about four years, although longer cycles are also believed to exist. Cycles vary widely in timing, pattern, and amplitude, both from one cycle to the next and from industry to industry. Major booms and depressions, however, affect nearly all economic activities.

Irregular fluctuations are the residual component in a time series after secular trend, cyclical, and seasonal movements have been accounted for. It is usually impossible, however, to separate cyclical and irregular fluctuations satisfactorily. The irregular factors may be "originating forces" (such as wars and acts of government) that influence business cycles or they may be miscellaneous unknown and unpredictable factors of a random nature.

Measures of business cycles are important in the study of past cyclical behavior, in forecasting business activity, and in planning stabilization policy. Cycles can be isolated by eliminating seasonality and perhaps trend by division or graphic adjustment and smoothing irregularities by a short-term moving average or freehand curve. The cyclical component remains as a residual. Sometimes only the seasonal adjustment is necessary. Computer programs such as Census II eliminate the calendar and seasonal components in successive steps and then smooth the residuals with a moving average of from one to six months, depending on the

⁸ See W. F. Butler and R. A. Kavesh, *How Business Economists Forecast* (Englewood Cliffs, New Jersey: Prentice-Hall, 1966); H. D. Wolfe, *Business Forecasting Methods* (New York: Holt, Rinehart & Winston, 1966); or the sources listed in J. B. Woy, *Business Trends and Forecasting* (New York: Gale Research, 1965) for further study.

irregularity of the data, to arrive at the trend-cycle component. Trend is left in, since it does not obscure the short-term cyclical pattern.

It is important to forecast the cyclical swings of business, particularly at turning points. A number of statistical forecasting methods are discussed: (1) various naïve methods in common use, (2) exponentially weighted moving averages, (3) lead and lag indicators, (4) diffusion indexes, (5) average duration of run, and (6) surveys of anticipations data. Statistical methods, however, must be supplemented by careful economic analysis to achieve an adequate forecast.

The statistical forecaster should be familiar with the materials in Chapters 2, 3, 14, and 19 to 24 of this book, as well as appropriate economics texts, as a basis for becoming adept in the strategic art of business forecasting.

PROBLEMS

1.
 - a) Select and plot a series of monthly data dominated by cyclical-irregular fluctuations rather than by secular or seasonal movements. The graph may be traced on a blank sheet placed over a chart in a current publication. Do not use textbook examples.
 - b) Describe its cyclical characteristics: Is the amplitude wide or narrow? How does the timing of the peaks and troughs compare with the timing of turning points in general business (Table 21-1)? What is the current phase of the cycle—expansion or contraction?
 - c) Describe the irregular movements: What was the behavior of this series during recent wars? What other major nonbusiness influences appear to have caused extended irregular movements? Are the month-to-month zigzag random forces marked or mild?
2.
 - a) List any peaks and troughs in general business that have occurred since February 1961 (from *Business Cycle Developments*, Appendix A) to update Table 21-1.
 - b) How did the National Bureau of Economic Research arrive at these "reference dates"?
 - c) Is there any evidence that expansion or contraction periods have changed in average length since World War II as compared with the period between the two world wars?
3. Which of the three purposes of measuring cycles is most important, in your opinion, for *a)* the business executive and *b)* the President's Council of Economic Advisers? Explain your choices.
4.
 - a) Outline both the graphic and the alternative arithmetic steps necessary to isolate the trend-cycle component of a time series.
 - b) Just how do these procedures eliminate seasonal and irregular influences? What traces of these elements are likely to remain in the trend-cycle residuals?

5. Cycles in monthly series are usually studied by examining data that are adjusted only for seasonal variation since secular trend rarely obscures short-term cycles and cyclical-irregular movements cannot be completely separated from each other. In your analysis of gasoline sales (Chapter 20, Problems 9 and 10), however, the cycles in the seasonally adjusted data (Problem 9 [e]) were obscured by secular trend and irregular elements. You therefore decide to eliminate these factors, as far as possible, in order to determine the nature of the cycle, if any, that may exist in this industry.
 - a) Trace the seasonally adjusted gasoline demand curve from Chapter 20, Problem 9 (e), onto another ratio chart and fit a straight trend line (since the trend is practically linear) by inspection, using the annual averages as guides.
 - b) Adjust the series for secular trend by laying off the vertical (not perpendicular) deviations above or below the trend line, with a divider or paper strip, around the horizontal line printed with "2" on the chart. Mark a "Percent of Trend" vertical scale with 50, 100, and 150 opposite the lines printed "1," "2," and "3," respectively. The curve is now adjusted for both seasonality and trend, so that it represents the estimated cyclical-irregular fluctuations in gasoline demand.
 - c) Draw a flexible freehand curve through the adjusted series to smooth out the month-to-month zigzags, but make it follow closely the short-term cyclical swings. This curve approximates the cycle itself (including extended irregular influences).
 - d) Describe the cyclical fluctuations, if any, in gasoline demand. In what months did cyclical peaks or troughs occur?
 6. If a computer program is available (e.g., Census II, Variant X-11), analyze Sears, Roebuck sales in Chapter 20, Table 20-2 (adding later sales as available) to:
 - a) Adjust for calendar and seasonal variation;
 - b) Smooth out irregularities with a short-term moving average, to isolate the trend-cycle component.
 - c) Also, interpret all results on your print-out sheet and hand it in with this sheet.
 7. Analyze the gasoline sales in Chapter 20, Problem 9, using the computer method outlined in Problem 6.
 8. Estimate the percent change in gross national product this year compared with last year, using any three of the four "naïve" methods of cyclical forecasting described in the text. Comment briefly on the validity of the results.
 9. Find an article on the use of exponentially weighted moving averages in sales forecasting and prepare a short report explaining this method (going beyond the textbook outline), together with its pros and cons.
 10. What is the present stage of the general business cycle—expansion or contraction? Is a turning point in prospect? Cite evidence supporting or modifying your view from:
-

- a) Lead and lag indicators.
 - b) Diffusion indexes.
 - c) A survey of anticipations data (e.g., businessmen's plans for new plant and equipment expenditures).
11. Select a leading indicator from *Business Cycle Developments* (as assigned) and:
- a) Explain on logical grounds why this indicator should lead general business at cyclical turning points.
 - b) Describe its performance and reliability in recent years as a barometer of business.
12. Prepare a critical review on the use of diffusion indexes (including average duration of run) as cyclical forecasting devices. Explanation should go beyond that in this text. See National Bureau of Economic Research publications, *Statistical Indicator Reports*, or *Business Cycle Developments*.
13. Select a survey of anticipations data, as assigned (see page 543, footnote 7 for sources), and report on its validity as a forecasting tool. Cite not only the original source but an outside critical study of its efficacy.

SELECTED READINGS

CROXTON, FREDERICK E., and COWDEN, DUDLEY J. *Practical Business Statistics*. 3d ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1960, chaps. 28 to 31, 34 and 38.

Explores numerous methods of isolating seasonal and cyclical fluctuations and trends, including the use of orthogonal polynomials and growth curves.

GRANGER, C. W. J., and HATANAKA, M. *Spectral Analysis of Economic Time Series*. Princeton, New Jersey: Princeton University Press, 1964.

A new technique which applies Fourier analysis to time series, and their interrelationships, using electronic computers.

DEWHURST, J. F. *America's Needs and Resources—A New Survey*. New York: The Twentieth Century Fund, 1955.

A detailed survey of trends in the economy, in some cases extending from 1850 to 2050.

MILLS, FREDERICK C. *Statistical Methods*. 3d ed. New York: Holt, Rinehart & Winston, 1955, chaps. 10 to 12 and Appendix F.

An authoritative treatment of time series analysis.

MITCHELL, WESLEY C. *What Happens during Business Cycles: A Progress Report*. New York: National Bureau of Economic Research, 1951.

Mitchell's works represent the most comprehensive statistical approach to business cycle analysis. See also his *Measuring Business Cycles*, with Arthur F. Burns, New York: National Bureau of Economic Research, Inc., 1946.

MOORE, GEOFFREY H. (ed.). *Business Cycle Indicators*. 2 vols., National Bureau of Economic Research. Princeton: Princeton University Press, 1961.

A comprehensive evaluation of cyclical measures of business in the United States and Canada, including lead and lag indicators, diffusion indexes, seasonal measurement, and the use of computers.

NETER, JOHN, and WASSERMAN, WILLIAM. *Fundamental Statistics for Business and Economics*. 3d ed. Boston: Allyn & Bacon, 1966.

Chapters 17 to 19 cover time series analysis for forecasting, planning, and control.

SHISKIN, JULIUS. *Signals of Recession and Recovery. An Experiment with Monthly Reporting*. New York: National Bureau of Economic Research, 1961.

Introduces the monthly indicators reported currently in *Business Cycle Developments*.

———, et al. *The X-11 Variant of the Census Method II Seasonal Adjustment Program*. U.S. Bureau of the Census, Technical Paper No. 15, November, 1965.

The latest census method, summarized in *Business Cycle Developments*, October, 1965.

22. SIMPLE CORRELATION AND REGRESSION

RELATIONSHIPS between variables are fundamental in science. The physical sciences have been highly successful in establishing functional relationships or "laws" connecting variables such as temperature and pressure of gas in a closed container, the distance of an object from the earth and the gravitational pull exerted upon it, and so on. The biological and social sciences have had to deal with more complicated situations in which there is less reason to expect exact relationships between variables. The statistical tools of correlation and regression analysis were developed to estimate the closeness with which two or more variables were associated and the average amount of change in one variable that was associated with a unit increase in the value of another variable. The term "regression" refers specifically to the measurement of this relationship. The more general term "correlation" includes regression analysis as well as certain other measures, such as the correlation coefficient. It is important to explore both the applications and limitations of these powerful tools of analysis in the study of economic relationships.

A preliminary step in studying the relationship between variables is to classify the data according to two or more characteristics in a cross-classification table, as outlined in Chapter 3. The present chapter describes more sophisticated methods for analyzing relationships. In particular, we shall explore the scatter diagram, curve fitting, estimation of population relationships from sample data, and the coefficient of correlation.

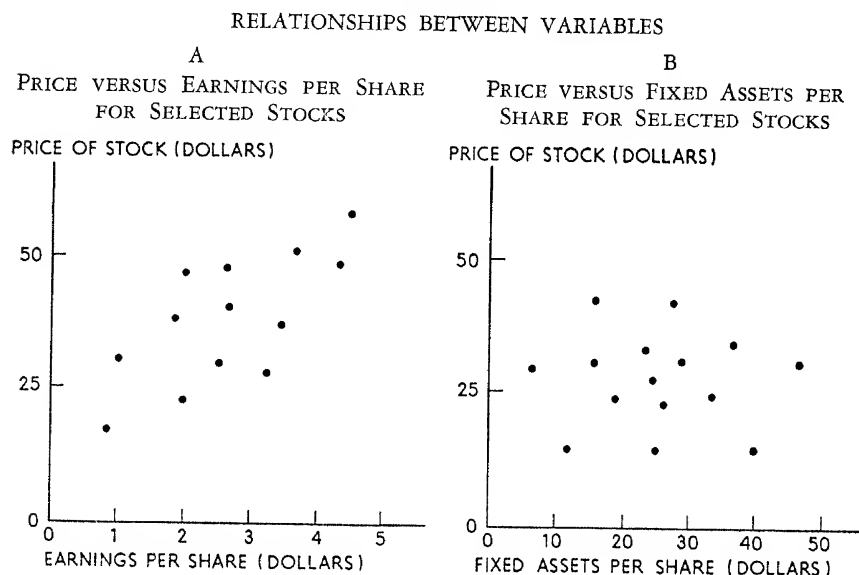
When only two variables are involved, the analysis is described as *simple* correlation or regression. *Multiple* correlation or regression refers to the analysis of three or more variables. This chapter is concerned

with simple (two-variable) relationships. The multiple variable case will be considered in Chapter 23.

SCATTER DIAGRAMS

A first step in analyzing the relationship between two variables is to plot the data on a chart called a *scatter diagram*. In Chart 22-1A, the prices of a group of stocks are related to the earnings per share. As is

Chart 22-1



evident from the diagram, stocks with higher earnings per share generally have higher prices. Thus, the two variables are related to, or *correlated* with, each other. Chart 22-1B illustrates a situation in which there is no apparent relationship between the price of stock and the fixed assets per share. We describe such variables as *uncorrelated* or as having *zero correlation*.

The correlation between two variables may be described as being *positive*, indicating that high values of one variable tend to be associated with high values of the other variable, and similarly with low values. For example, in Chart 22-2A, families with higher incomes tend to spend more for housing than families with lower incomes, so the plotted points move upward to the right. When high values of one variable occur with low values of the other, the variables are inversely or *negatively* correlated. Thus, in Chart 22-2B, a larger crop of pigs means a lower price, so the points move downward from left to right.

Chart 22-2

POSITIVE AND NEGATIVE CORRELATION

A
FAMILY INCOME VERSUS EXPENDITURES FOR HOUSING FOR SELECTED FAMILIES

EXPENDITURES FOR HOUSING
(THOUSANDS OF DOLLARS)



B
MILLIONS OF PIGS RAISED
VERSUS PRICE OF HOGS FOR
SELECTED YEARS

PRICE OF HOGS (DOLLARS)

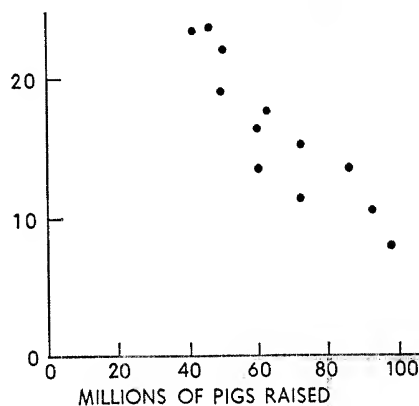
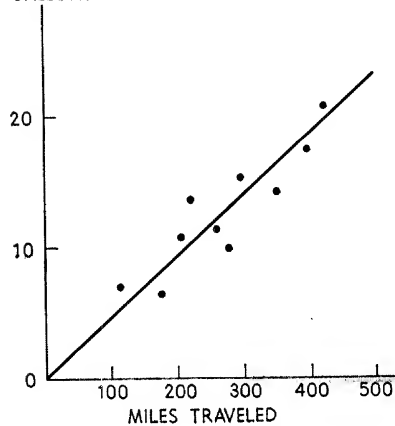


Chart 22-3

LINEAR AND CURVILINEAR CORRELATION

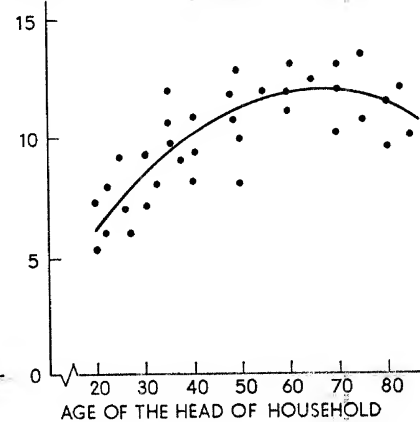
A
GALLONS OF GAS VERSUS MILES
TRAVELED FOR SELECTED
TRIPS

GALLONS OF GAS USED



B
FAMILY INCOME VERSUS AGE OF
THE HEAD OF THE HOUSEHOLD
FOR SELECTED FAMILIES

FAMILY INCOME
(THOUSANDS OF DOLLARS)



If the plotted points on a scatter diagram generally follow a straight line, we say that there is a *linear* relationship between the two variables. This is true of Chart 22-3A, where each hundred miles of travel on a trip requires about the same number of gallons of gasoline. Note that the straight line is a good fit to the plotted points. If a curved line gives a better fit, the correlation is said to be *curvilinear*. In Chart 22-3B, income at first rises with the age of the head of household, then levels off, and finally falls as retirement age is reached. The curve, as drawn, follows the data more closely than would a straight line.

REGRESSION ANALYSIS

In the previous section, we introduced the scatter diagram as a graphic means of presenting the relationship between two variables. In most business and economic situations, however, we wish to use one of the variables to *predict* or *control* the other variable. Hence, we need techniques for prediction and for measuring the error in our predictions. These techniques are called *regression analysis*.

Curve Fitting

The first step is to express the relationship between the two variables as a line or mathematical equation. The variable to be predicted is designated as Y , the *dependent* variable. The other variable, X , is the *independent* or *predicting* variable. The dependent variable is then expressed as some function of the independent variable; i.e., $Y = f(X)$. This regression function is similar to the trend function discussed in Chapter 19, except that some variable other than time is used as the independent variable.

The simplest functional form is the straight line. The formula for a straight line is $Y_c = a + bX$, where Y_c is the computed value of Y (i.e., the value on the line for a given value of X). The constant a is the value of Y_c at the Y axis when $X = 0$, and b is the increase in Y_c for each unit increase in X . The value of b is therefore the slope of the line. When a straight line is used to relate two variables, the regression equation is said to be *linear*. The slope b is then termed the *regression coefficient*. This chapter is primarily concerned with linear relationships. Fortunately, the straight line is adequate for relating variables in many business and economic situations. If a straight line is not a good fit in representing the relationship between the variables, the graphic method described below or the mathematical techniques suggested in Chapter 24 should be employed.

An example will serve to introduce the concepts and techniques of regression analysis. The personnel manager in an electronic manufac-

turing company devises a manual dexterity test for job applicants to predict their production rating in the assembly department. In order to

Table 22-1
SCORES ON MANUAL DEXTERITY TEST AND PRODUCTION
RATINGS FOR 20 WORKERS

Worker	Test Score X	Production Rating Y
A.....	53	45
B.....	36	43
C.....	88	89
D.....	84	79
E.....	86	84
F.....	64	66
G.....	45	49
H.....	48	48
I.....	39	43
J.....	67	76
K.....	54	59
L.....	73	77
M.....	65	56
N.....	29	28
O.....	52	51
P.....	22	27
Q.....	76	76
R.....	32	34
S.....	51	60
T.....	37	32

do this, he selects a random sample of 20 applicants. They are given the test and later assigned a production rating. It is a common practice to administer an aptitude test to applicants for jobs, especially for types of jobs which require similar skills and for which objective measures of success can be obtained later.

The results are shown in Table 22-1 and Chart 22-4, where each dot represents one employee. The test score is the independent variable. There seems to be a fairly close linear relationship, with the dots clustered along a straight line, and with no extreme deviations.

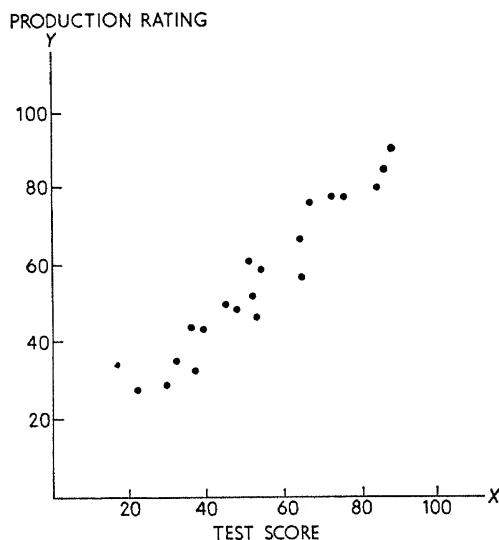
Our object is to find the values of a and b in the straight line, $Y_e = a + bX$, which will predict production rating (Y_e) for any applicant's test score (X).

Since the points in Chart 22-4 are somewhat scattered, we cannot predict production ratings (Y) exactly. For any given test score, the predicted value Y_e is roughly the average of the production ratings (Y 's) with the given test score. Thus, the regression line is often called

the *line of average relationship*, indicating that it is a plot of the average values of Y for different values of X . The deviations of the actual ratings from the averages ($Y - Y_c$) are due to various personal differences and flaws in the test as a predictive device.

Chart 22-4

SCATTER DIAGRAM SHOWING RELATIONSHIP
BETWEEN TEST SCORES AND PRODUCTION
RATINGS FOR 20 WORKERS



Two methods of fitting a straight line are described below: the graphic "freehand" and the method of least squares. The graphic method has the advantages of being simple and flexible in shape as well as permitting the skilled analyst to minimize the influence of extreme cases and otherwise follow the logical implications of the data. On the other hand, the method of least squares has the advantage of being objective and precise and is easily adapted to large-scale machine computation. The graphic method is often used as a preliminary sketch to determine the general nature of the relationship upon which the appropriate mathematical curve is fitted.

Graphic Method. The steps to be followed in the graphic method may be summarized as follows. Draw the line through the plotted points by inspection so that the *vertical* deviations of the dots above and below the line are exactly equal for the series as a whole and are approximately equal for each major segment of the plotted data. These deviations may be marked off accumulatively on the edge of a strip of paper, one above the other, for comparison.

When the dots in the scatter diagram are numerous or widely scattered, the average values of groups of data should be plotted to serve as objective guide points in drawing the regression line or curve. First, divide the data into several groups according to values of X , each group having about the same number of items. Using too many groups will lead to a zigzag pattern in the group averages; using too few groups will make the averages insensitive as guides to the shape of the estimating line.

Second, take the mean of the X and Y values in each group, and plot this group average on the scatter diagram.

Third, draw a smooth line or curve (using a transparent ruler or French curve) between the plotted averages, so that the vertical deviations of the averages above the line *exactly* equal those below the line over the whole range, and are approximately equal for each of several broad segments along the line. In particular, if the group averages follow a fairly straight line (except for zigzags), plot the overall mean (\bar{X} , \bar{Y}) and draw a straight line through this point at such a slope as to equalize approximately the vertical deviations of the group averages on the left of this point and those on the right separately. A curve should be drawn only if the group averages follow an unmistakable curve which is supported by economic logic.

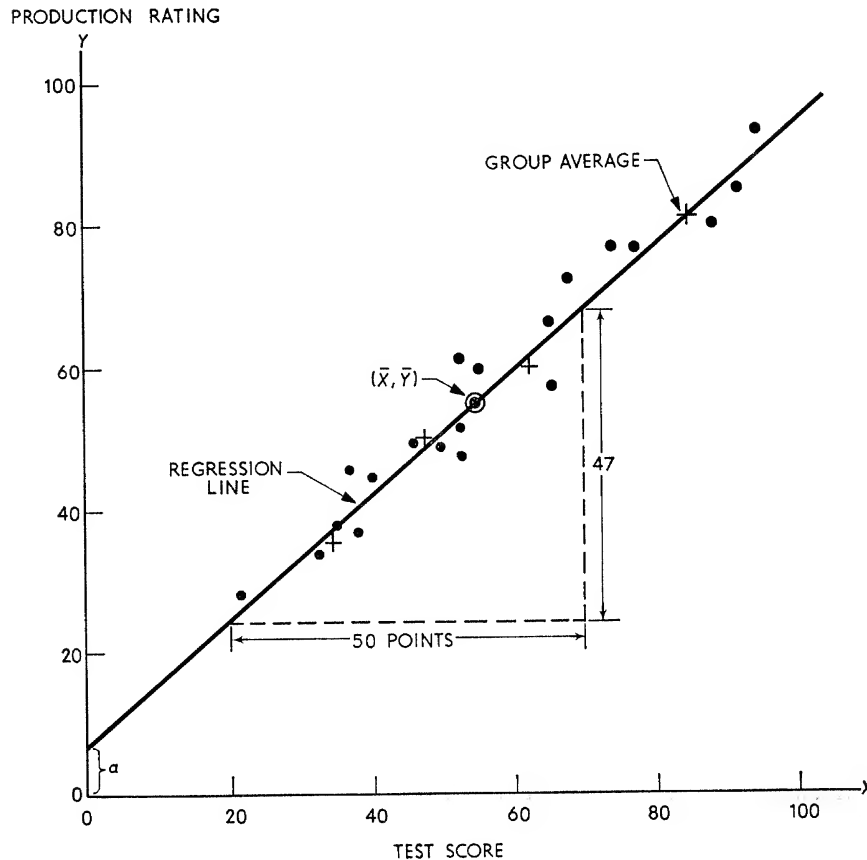
Most beginners have a tendency to draw graphic regression curves too steep because they judge goodness of fit by the shortest (or perpendicular) distance from the point to the line rather than by the vertical distance (the direction in which the dependent variable Y is measured) from the point to the line. Curvature of the regression aggravates this tendency, especially in the part of the chart where the regression is steepest. The use of group averages reduces this error.

In our example of test scores and production ratings, the steps outlined above have been performed on Chart 22-5. Crosses indicate averages of four groups of points, and the overall average (\bar{X} , \bar{Y}) is circled. A straight line is drawn through the overall average and as close to the group averages as possible. The values of a and b for the regression line are estimated from the chart. The line crosses the Y axis (when $X = 0$) at approximately 4.0. Thus, the intercept a is 4.0. Over 50 points of test score (from 20 to 70), the value of Y_c increases from 23 to 70, a difference of 47 units on the production rating scale. Thus, the slope is estimated to be $47/50 = 0.94$. This is the regression coefficient b . The graphic estimate of the regression line can now be written as

$$Y_c = 4.0 + 0.94X$$

Chart 22-5

GRAPHIC METHOD OF ESTIMATING PRODUCTION RATINGS FROM TEST SCORES
FOR 20 WORKERS



The Method of Least Squares. A straight line fitted by least squares has the following characteristics:

1. It gives the best fit to the data in the sense that it makes the sum of the squared deviations from the line, $\Sigma(Y - Y_o)^2$, smaller than they would be from any other straight line. This property accounts for the name "least squares."
2. The deviations above the line equal those below the line, on the average. This means that the total of the positive and negative deviations is zero, or $\Sigma(Y - Y_o) = 0$.
3. The straight line goes through the overall mean of the data (\bar{X}, \bar{Y}) .
4. When the data represent a sample from a larger population, the

least squares line is a "best" estimate of the population regression line. This property will be discussed in more detail later.

It is important to stress that the deviations $(Y - Y_c)$ are measured vertically (i.e., along the Y axis). The deviations are *not* perpendicular to the regression line.

For the least squares line the values of a and b in the equation $Y_c = a + bX$ are found by solving the two normal equations

$$\begin{aligned}\Sigma Y &= na + b\Sigma X \\ \Sigma XY &= a\Sigma X + b\Sigma X^2\end{aligned}$$

where n is the number of pairs of items in a sample.

The computations can be simplified in most problems by measuring both X and Y as deviations from their means \bar{X} and \bar{Y} . These deviations are designated by the small letters x and y , where $x = X - \bar{X}$ and $y = Y - \bar{Y}$. It is not necessary, however, to subtract the mean from each value of X and Y . A simpler procedure is as follows:

1. Compute the product XY , and calculate or look up the squares X^2 and Y^2 in a table, for each original pair of observations.
2. Sum these columns. (Steps 1 and 2 can be combined in a single operation on a calculating machine.)
3. Subtract from each sum the *mean times the sum* of the respective variables to get the adjusted sums of the x 's and y 's expressed as deviations from their means. That is,¹

Sum	ΣXY	ΣX^2	ΣY^2
Less mean times sum	$-\bar{X}\Sigma Y$	$-\bar{X}\Sigma X$	$-\bar{Y}\Sigma Y$
Equals adjusted sum	$= \Sigma xy$	$= \Sigma x^2$	$= \Sigma y^2$

The sum of the deviations around the means, Σx and Σy , must equal zero, so they drop out of the two normal equations above, which reduce to

$$\begin{aligned}b &= \frac{\Sigma xy}{\Sigma x^2} \\ a &= \bar{Y} - b\bar{X}\end{aligned}$$

where b derives from the second normal equation when $\Sigma x = 0$, and a is obtained by solving the first equation intact to express it in the original units.

For our illustration of test scores and production ratings, the calcula-

¹ Note that $\Sigma x^2 = \Sigma (X - \bar{X})^2 = \Sigma (\bar{X}^2 - 2\bar{X}X + X^2) = \Sigma X^2 - 2\bar{X}\Sigma X + n\bar{X}^2$. But since $n\bar{X} = \Sigma X$, we have $\Sigma x^2 = \Sigma X^2 - 2\bar{X}\Sigma X + (\Sigma X)\bar{X} = \Sigma X^2 - \bar{X}\Sigma X$. The formulas for Σy^2 and Σxy can be derived in a similar fashion.

Table 22-2

CORRELATION BETWEEN SCORES ON MANUAL DEXTERITY TEST
AND PRODUCTION RATINGS FOR 20 WORKERS

Worker	Test Score X	Production Rating Y	XY	X ²	Y ²
A	53	45	2,385	2,809	2,025
B	36	43	1,548	1,296	1,849
C	88	89	7,832	7,744	7,921
D	84	79	6,636	7,056	6,241
E	86	84	7,224	7,396	7,056
F	64	66	4,224	4,096	4,356
G	45	49	2,205	2,025	2,401
H	48	48	2,304	2,304	2,304
I	39	43	1,677	1,521	1,849
J	67	76	5,092	4,489	5,776
K	54	59	3,186	2,916	3,481
L	73	77	5,621	5,329	5,929
M	65	56	3,640	4,225	3,136
N	29	28	812	841	784
O	52	51	2,652	2,704	2,601
P	22	27	594	484	729
Q	76	76	5,776	5,776	5,776
R	32	34	1,088	1,024	1,156
S	51	60	3,060	2,601	3,600
T	37	32	1,184	1,369	1,024
Sum	1,101	1,122	68,740	68,005	69,994
Mean	55.05	56.10			
Less mean times sum.....			-61,766	-60,610	-62,944
Equals adjusted sum.....			6,974	7,395	7,050
This is.....			Σxy	Σx^2	Σy^2

tions are shown in Table 22-2. We compute XY , X^2 , and Y^2 for each worker, sum these, and subtract the respective mean times the sum (shown in the box under X and Y) to find Σxy , Σx^2 , and Σy^2 . Then

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{6,974}{7,395} = 0.943$$

$$a = \bar{Y} - b\bar{X} = 56.10 - 0.943(55.05) = 4.2$$

Hence, the regression line is

$$Y_c = 4.2 + 0.943X$$

If a job applicant from the same population received a test score of 40, therefore, his production rating could be estimated as

$$Y_c = 4.2 + 0.943(40) = 42$$

Alternatively, this value might be read graphically from Chart 22-6 (dotted lines).

The Standard Error of Estimate

The usefulness of the regression line for purposes of prediction and control depends on the extent of the scatter of the observations about it. If the observed values of Y vary widely about the line, estimates of Y based on this line will not be very accurate. On the other hand, if the observed values of Y lie quite close to the line, the estimates based on this line may be very good. The measure of the scatter of the actual observations about the regression line is called the *standard error of estimate*. The standard error of estimate for the population may be estimated from a sample in linear regression as follows:

$$S_{YX} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - 2}}$$

where n is the size of the sample.²

The value $\Sigma(Y - Y_c)^2$ can be obtained graphically by reading off the *vertical* (not perpendicular) deviation of each point (Y) from the regression line (Y_c) on the Y scale, squaring each deviation, and summing these squares. The value Y_c can also be computed from the regression equation for each given value of X , to find $\Sigma(Y - Y_c)^2$.

When a straight line regression has been fitted by least squares, however, it is usually simpler to compute the standard error of estimate by the following formula:

$$S_{YX} = \sqrt{\frac{\Sigma y^2 - b\Sigma xy}{n - 2}}$$

² The standard error of estimate for the sample itself is $\sqrt{\Sigma(Y - Y_c)^2/n}$. The use of $n - 2$ adjusts for sample bias. This number represents the degrees of freedom around the regression line, just as $n - 1$ was used as the number of degrees of freedom around the mean in computing the standard deviation. Whereas the selection of the sample mean as a point from which to measure $Y - \bar{Y}$ uses up only one degree of freedom, the selection of a straight regression line as a base from which to measure the scatter uses up two degrees of freedom: one in requiring that the line pass through the point of means (\bar{X}, \bar{Y}) and the other in determining the slope of the regression line. In general, the adjustment is $n - k$, where k is the number of constants in the regression equation.

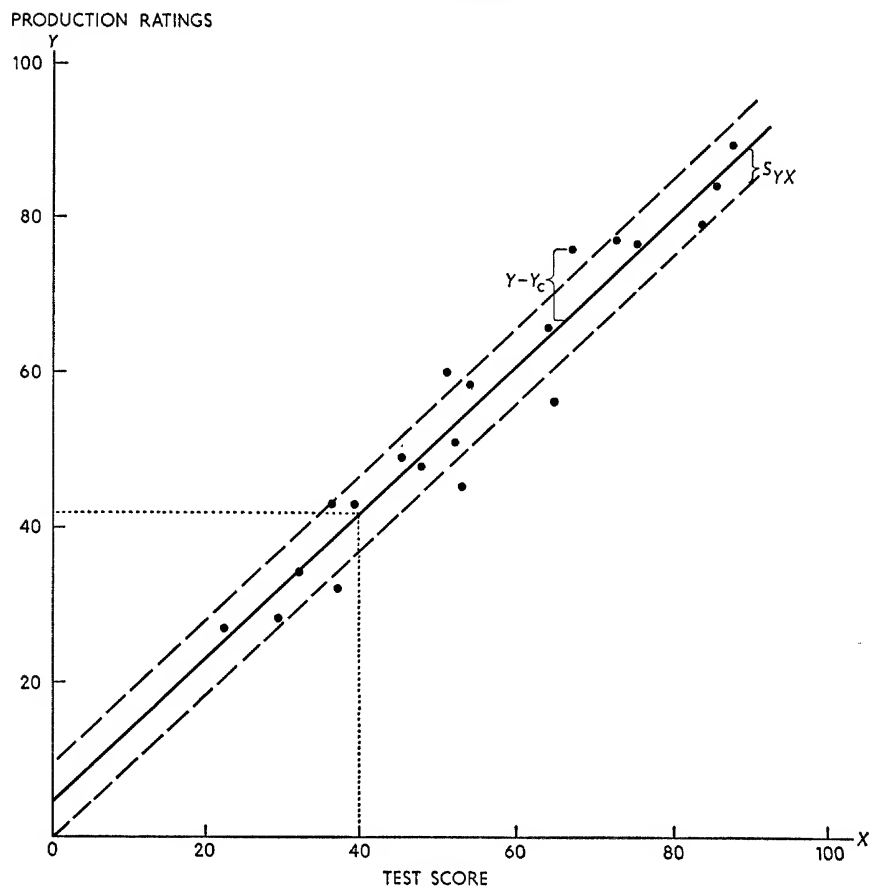
Thus, in our example of test scores and production ratings (Table 22-2):

$$\begin{aligned}
 S_{YX} &= \sqrt{\frac{\Sigma y^2 - b\Sigma xy}{n - 2}} \\
 &= \sqrt{\frac{7,050 - 0.943(6,974)}{20 - 2}} \\
 &= 5.13
 \end{aligned}$$

The standard error of estimate has been laid off in Chart 22-6 above and below the regression line (see dashed lines). If the points are

Chart 22-6

REGRESSION LINE FITTED BY LEAST SQUARES
AND STANDARD ERROR OF ESTIMATE
TEST SCORES AND PRODUCTION RATINGS
OF 20 WORKERS



Source: Table 22-2.

scattered at random about the regression line (i.e., if $z = Y - Y_c$ follows a nearly normal distribution), then approximately two thirds of the points should lie within this band. Hence, management could predict that an applicant who scored 40 on the test would achieve a production rating of 42 ± 5 , or between 37 and 47, with two chances out of three of being correct. This standard error can also be compared with the standard error of estimate based on the use of alternative aptitude tests as predictors, that is, mechanical aptitude, mathematical ability, etc. In this way, it is possible to compare the performance of various alternative tests as predictors of success on a given type of job.

SAMPLING AND REGRESSION ANALYSIS

Up to this point we have considered the regression line and standard error of estimate merely as *descriptions* of the average relationship between two variables and of the goodness of fit.

However, we are not usually interested in regression results solely as a description of a particular sample. Almost without exception we are looking for a relationship that will enable us to control or predict new values of the dependent variable within limits of accuracy estimated from the original set of data.

Thus, regression analysis of business and economic statistics must be approached from the standpoint of (statistical) inference from a particular sample to a "parent population" which includes the given sample and also such future or additional observations as we wish to control or predict. Both the given sample which we analyze and the actual future values or "drawings" we attempt to control or predict represent only a fraction of all of the possible values that might conceivably be drawn from the population in question. The application of statistical inference to regression analysis leads to the discovery and verification of relationships between variables. This is one of the most challenging and basic problems of scientific research.

The regression line for a sample is only one of a family of regression lines for different samples that might be drawn from the same population. That is, regression measures are subject to sampling error. Nevertheless, we can estimate within what limits the "true" regression line in the population is likely to fall. The theory of estimating population parameters from sample statistics was introduced in Chapters 11 and 12. We can now apply this theory in making statistical inferences about the true values of regression and correlation parameters.³

³ See M. Ezekiel and K. A. Fox, *Methods of Correlation and Regression Analysis* (3d ed.; New York: John Wiley, 1959), chaps. 17 and 19, for a more complete discussion of this topic.

Basic Assumptions

In order to make valid inferences from sample data about population relationships, certain assumptions must be satisfied.

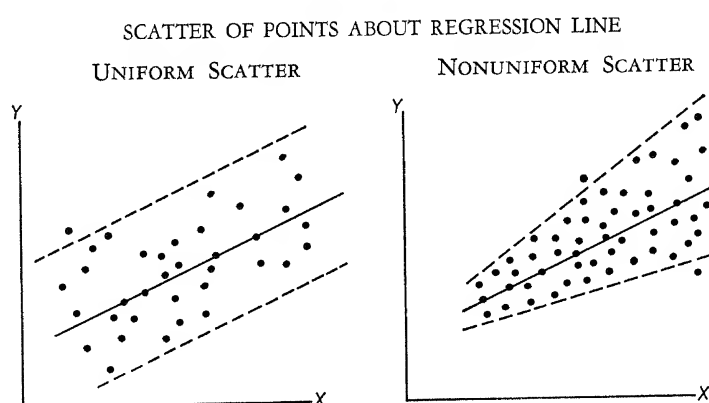
Assumption 1. When we fit a straight line to sample data to estimate the true or population relationship, the latter must also be linear. (The curvilinear case is described in Chapter 24.) This underlying relationship may be expressed in the form

$$Y = A + BX + z$$

where A and B are the true (but unknown) parameters of the regression line, and z is the deviation of an actual value of Y from the true regression line. That is, $z = Y - Y_c$. (The average or expected value of z is zero.) This is the assumption of linearity.

Assumption 2. The standard deviation of the z 's is the same for

Chart 22-7



all values of X . This means that there is a uniform scatter or dispersion of points about the regression line. This property is called *homoscedasticity*.⁴ Examples illustrating when this assumption is valid and when it is invalid are shown in Chart 22-7.

Assumption 3. The z 's are *independent* of each other. This means that the deviation of one point about the line (its z value) is not related to the deviation of any other point. This assumption of independence is *not* valid for most time series data. Time series move in cycles rather than randomly about the trend, so that adjoining values (e.g., in two boom years) are closely related. Independent and dependent data are

⁴ When the scatter is not uniform, it is sometimes possible to make the assumption valid by means of a transformation (e.g., convert Y to $\log Y$).

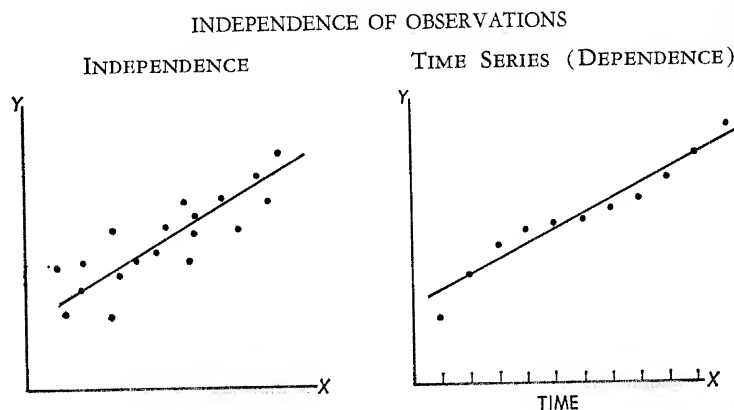
illustrated in Chart 22-8. Chapter 24 includes some techniques for least squares regression of time series when the independence assumption is not valid.

Assumption 4. The distribution of the points above and below the regression line follows a roughly normal curve. This means that the z values are normally distributed.⁵

When these four assumptions are satisfied, the linear regression coefficients and standard error of estimate computed from a sample are efficient, linear unbiased estimators of the true population values.

In addition to these general assumptions, it is important to distin-

Chart 22-8



guish between two cases, called the correlation model and the regression model.

Correlation Model. In the correlation model, both X and Y are considered to be random samples drawn from a normal population.⁶

⁵ The necessity of Assumption 4 in determining the validity of regression measures depends upon the size of the sample.

For small samples, normality of the z values is not necessary if one wishes only to estimate the values a and b of the regression line. However, the assumption is necessary for the valid use of the standard error measures, such as s_b and s_{Y_c} , considered below. The normality assumption is also necessary in order to make probability statements using the standard error of estimates S_{YX} and the standard error of forecast s_{Y-Y} (below).

For large samples, the normality of the z values is not necessary to make valid inferences about the regression line (i.e., to make inferences about a and b using the standard error measures s_b and s_{Y_c}). The central limit theorem enables us to make such inferences despite the non-normality of the z values. However, normality is necessary to make probability statements using S_{YX} and s_{Y-Y} .

See A. M. Mood and F. A. Graybill, *Introduction to the Theory of Statistics* (2d ed., New York: McGraw-Hill, 1963), Chapter 13, for more detail on the properties of these estimators.

⁶ More specifically, the data pairs (X, Y) should represent a random sample from a population that is normal with respect to both variables.

The sample values are thus independent of each other and are normally distributed about their respective means. If this condition is met, together with the four general assumptions listed above, all correlation and regression measures in this chapter may be considered valid.

Regression Model. In the regression model, Y is a random variable, but X is fixed or predetermined at specific values. This is often true of controlled experiments. For example, in measuring the effects of various amounts of fertilizer upon corn yields, the X values may be determined as 0, 40, 80, and 120 pounds of nitrogen, respectively, in four groups of plots. In this case, regression analysis is valid only for other samples or a population in which the X values are selected in exactly the same manner as in the original sample, for example, for plots of 0, 40, 80 and 120 pounds of fertilizer drawn with the same frequency as in this sample. The coefficient of correlation (described below) is generally not valid in the regression model.

We now turn to the problem of measuring the sampling error associated with the estimates a and b and the statistical inferences that can be drawn based upon these estimates.

Sampling Error of the Regression Coefficient

An inference about a regression coefficient can be made either as a test of significance or as a confidence interval, just as in the case of the mean or a proportion. Either type of inference depends on the standard error of the regression coefficient, as described below.

Testing the Significance of a Relationship. In the first place, it might be useful to know if there is *any* significant relationship between the variables X and Y . Some particular sample may indicate a relationship, even when none exists, by pure chance. If there is no relationship, then the slope B of the true regression line would be zero. This, then, is set up as the hypothesis, that is, $B = 0$. If the sample value b is significantly different from zero, we reject the hypothesis and assert that there is a definite relationship between the variables. To do all this, we compute the *standard error of the regression coefficient*. This is

$$s_b = \frac{S_{YX}}{\sqrt{\sum x^2}}$$

Here, S_{YX} is the sample standard error of estimate, and $\sum x^2$ describes the dispersion of X values around their mean. The value s_b is a measure of the amount of sampling error in b , just as $s_{\bar{X}}$ was a measure of the sampling error in the mean \bar{X} .

In the production rating example (Table 22-2):

$$s_b = \frac{5.13}{\sqrt{7,395}} = 0.060$$

The procedure for deciding whether a positive relationship exists between production ratings and test scores may be set forth as follows:

Null hypothesis: $B = 0$ (No relationship between production ratings and test scores)

Alternative hypothesis: $B > 0$ (Production rating increases as test score increases)

The value of b is 0.943. If the null hypothesis is true, $B = 0$ and b is 0.943 units from B . In terms of its standard error, this is $0.943/s_b = 0.943/0.060 = 16$. Thus b is 16 standard errors from $B = 0$.

If this analysis were based upon a large sample, the one-tailed probability associated with any given deviation could be found from the table of areas under the normal curve in Appendix D. For small samples such as this one (with $n \leq 30$), the t distribution in Appendix J must be used with $n - 2$ degrees of freedom. In either case, a deviation of more than three standard errors is highly significant (except for very small samples). The chance is negligible, therefore, that a deviation as large as 16 standard errors could occur by chance. Hence, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between the variables.

Confidence Intervals

A useful way to express the amount of sampling error in sample statistics is by means of confidence intervals. Confidence intervals will be illustrated here for

1. The regression coefficient or slope of the population regression line (B).
2. The population value for any point on the regression line.
3. An individual forecast.

The 95 percent confidence interval will be illustrated here, but any other degree of confidence may be chosen instead, by reference to Appendix D or J.

The Regression Coefficient. The 95 percent confidence interval for the regression coefficient in a large sample is

$$b \pm 1.96s_b \quad (\text{Appendix D})$$

In the production rating example, however, with $n = 20$, we look up Appendix J with $n - 2 = 18$ degrees of freedom and $P = 0.05$ to find the confidence interval

$$\begin{aligned} & b \pm 2.10s_b \\ \text{This is } & 0.943 \pm 2.10(.060) \\ & = 0.943 \pm 0.126 \end{aligned}$$

The manufacturer therefore could make the statement that B is between 0.817 and 1.069, with a probability of 0.95 that this statement is correct.

The Regression Line. A regression line obtained from a sample will vary from the true regression line not only in its slope but also in its elevation. The average height of the line is best determined by the estimated mean of the Y values, \bar{Y} . The standard error of the mean is

$$s_{\bar{Y}} = \frac{S_{YX}}{\sqrt{n}}$$

The standard error for any point Y_o on the regression line may now be determined from the equations for $s_{\bar{Y}}$ and s_b . We can express the regression equation in the form $Y_o = \bar{Y} + bx$. The standard error of Y_o for any value of x (the deviation from the mean) will then include the standard errors of both \bar{Y} and $b(x)$. Standard errors, like standard deviations, may be summed by adding their squares. The standard error of Y_o for any value of x , therefore, is derived as follows:

$$\begin{aligned} s_{Y_o}^2 &= s_{\bar{Y}}^2 + (s_b x)^2 \\ &= \frac{S_{YX}^2}{n} + \frac{S_{YX}^2 x^2}{\Sigma x^2} \end{aligned}$$

The standard error of a point on the regression line is therefore

$$s_{Y_o} = S_{YX} \sqrt{\frac{1}{n} + \frac{x^2}{\Sigma x^2}} \quad \text{for each value of } x = X - \bar{X}$$

In the production rating example, $S_{YX} = 5.13$, $n = 20$, and $\Sigma x^2 = 7,395$ (Table 22-2). Therefore,

$$s_{Y_o} = 5.13 \sqrt{\frac{1}{20} + \frac{x^2}{7,395}}$$

The standard error of the regression line is smallest at \bar{X} , when $x = 0$, and increases in either direction. Its values are shown in Table 22-3, column 4, for selected values of the test score X .

The 95 percent confidence interval for the regression line, when

$n = 20$, is $Y_c \pm 2.10s_{Y_c}$. This is shown by the dashed lines in Chart 22-9. The chances are 95 out of 100, therefore, that the true regression line for the population falls within these limits.

An Individual Forecast. It is often important to find within what limits a *new observation* may be expected to lie. For example, the regression line in Chart 22-6 was used to forecast the production rating for a new applicant who received a test score of 40. The estimated rating was 42 ± 5 , where 5 was the standard error of estimate. This error, however, did not take into account the sampling error in the regression line itself.

Table 22-3

STANDARD ERROR OF REGRESSION LINE
AND STANDARD ERROR OF AN INDIVIDUAL FORECAST
TEST SCORES AND PRODUCTION RATINGS OF 20 WORKERS

SELECTED VALUE OF X (1)	DEVIATION FROM MEAN, x (2)	x^2 7,395 (3)	STANDARD ERROR OF	
			Regression Line, s_{Y_c} (4)	Forecast s_{Y-Y_c} (5)
15	-40	0.2164	2.65	5.77
35	-20	0.0541	1.65	5.39
55	0	0	1.15	5.26
75	20	0.0541	1.65	5.39
95	40	0.2164	2.65	5.77

Note: For 95 percent confidence intervals multiply columns 4 and 5 by 2.10.
Source: Table 22-2.

The *standard error of forecast* (s_{Y-Y_c}) is a measure of the total sampling error for any new observation. It is obtained by combining the standard error of estimate (S_{YX}) and the standard error of the regression line (s_{Y_c}). The standard errors must be squared and added, as follows:

$$s_{Y-Y_c}^2 = S_{YX}^2 + s_{Y_c}^2$$

Substituting the value of s_{Y_c} found above, the formula for the standard error of forecast becomes

$$s_{Y-Y_c} = S_{YX} \sqrt{1 + \frac{1}{n} + \frac{x^2}{\sum x^2}} \text{ for each value of } x = X - \bar{X}$$

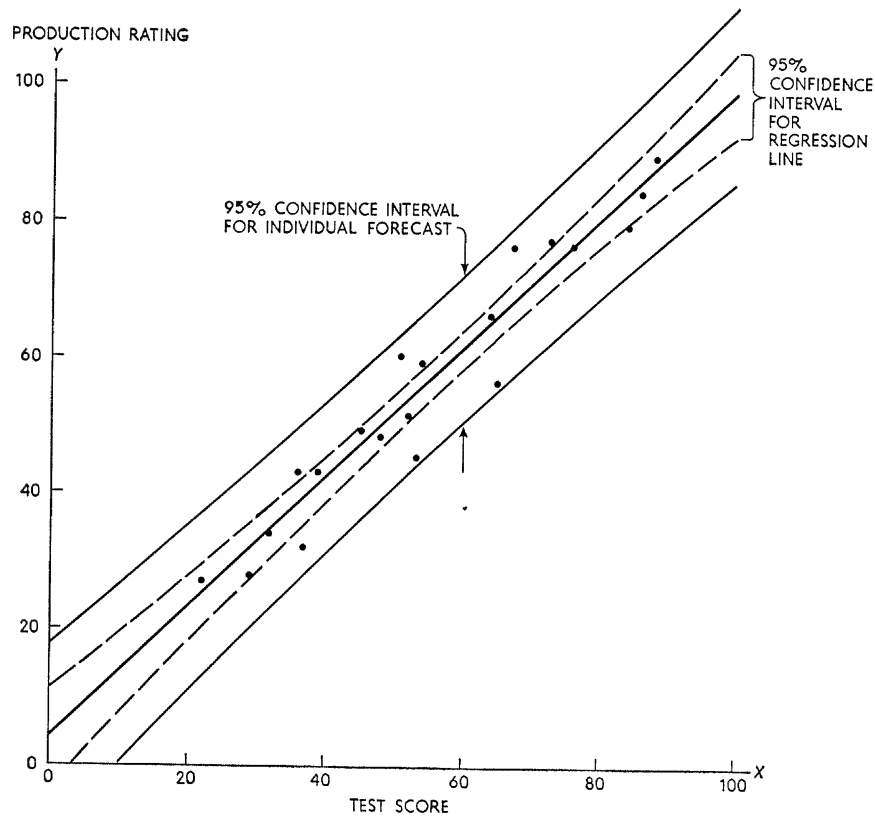
This formula simply adds 1 under the radical to the formula for the standard error of the regression line.

In the production rating case, the standard error of forecast is

$$s_{Y-Y_c} = 5.13 \sqrt{1 + \frac{1}{20} + \frac{x^2}{7,395}}$$

Chart 22-9

CONFIDENCE INTERVALS FOR REGRESSION LINE
AND INDIVIDUAL FORECAST
TEST SCORES AND PRODUCTION RATINGS
OF 20 WORKERS



SOURCE: Tables 22-2 and 22-3.

The forecast errors for five selected test scores (X) are given in Table 22-3, column 5.

If the calculations for the forecast error are based upon a large sample, and if the values are approximately normally distributed about the regression line, then the chances are about 95 percent that a new observation drawn from the same population will be within 1.96 forecast errors on either side of Y_c . That is to say, the 95 percent confidence interval for a new observation (Y) is $Y_c \pm 1.96s_{Y-Y_c}$.

In the present example, however, with sample size only 20, the 95 percent confidence interval for a new observation is $Y \pm 2.10s_{Y-Y_c}$. This interval is shown as the wide band in Chart 22-9. The chances are

95 out of 100, therefore, that a new applicant will achieve a production rating within these limits.

Certain characteristics of Chart 22-9 should be carefully observed. The boundaries of the confidence intervals are curved. The further the X values get from their arithmetic mean, the greater the width of the confidence intervals. This fact points up the danger of extrapolating for values of X that are a considerable distance from \bar{X} .

The forecast error is useful not only for *prediction* but also for *control*. If an observation falls outside the confidence limits, this indicates that it is very likely "out of control" and should be investigated. As a control chart, Chart 22-9 serves much the same purpose as the statistical quality control charts described in Chapter 25. In the present example, management can not only *predict* that an applicant with test score of 40 will achieve a production rating between 31 and 53 (with probability 95 percent), but they can use these points as *control* limits. If the applicant's actual production rating falls outside these limits, the chart warns the supervisor to investigate. If the employee's production is below 31, it may be possible to identify and remedy the cause of this deficiency; if it is above 53, the factors accounting for this superior performance should also be identified, either as a basis of rewarding the employee or improving work practices generally.

COEFFICIENT OF CORRELATION

The coefficient of correlation (r) is a relative measure of the relationship between two variables. It varies from zero (no correlation) to ± 1 (perfect correlation). The sign of r is the same as that of b in the regression equation. Thus, if $r = -1$, all dots are on a regression line sloping down to the right.

More specifically, the correlation coefficient may be defined as a measure of the extent to which the independent variable accounts for the variability in the dependent variable. This concept is illustrated in Chart 22-10. Note that the total deviation of the dependent variable Y from its mean \bar{Y} can be broken into two parts: the deviation of the value on the line from the mean ($Y_c - \bar{Y}$), which is *explained* by the given value of X , and the deviation of Y from the regression line ($Y - Y_c = z$), which is *not* explained by X . That is, $(Y - \bar{Y}) = (Y_c - \bar{Y}) + (Y - Y_c)$.

Since the two parts are independent, the total variance of Y may be expressed as the sum of the variances of the two parts:

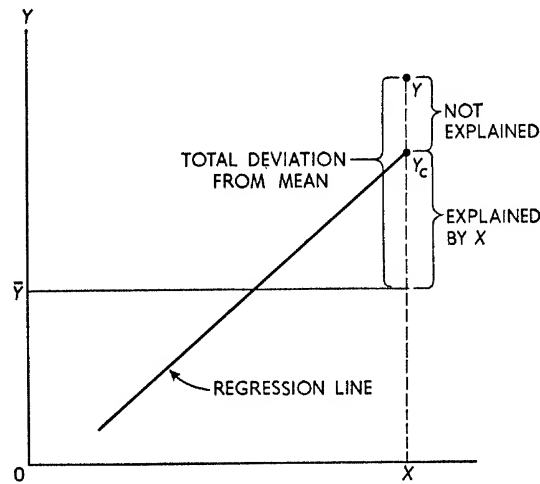
$$s_Y^2 = s_{Y_c - \bar{Y}}^2 + s_{Y - Y_c}^2$$

The standard error of estimate (S_{YX}^2) measures the deviations of the points about the line. It thus represents the variance in Y that remains (i.e., the *unexplained* variance) after the regression line has been fitted to the data. The term $s_{Y_c - \bar{Y}}^2$ is the variance of points on the regression line around the mean value \bar{Y} (or the variance *explained* by the regression line).

By expressing the explained variance as a ratio of the total variance

Chart 22-10

BASIC MEASURES FOR CORRELATION COEFFICIENTS



of Y , we obtain the *square* of the correlation coefficient, called the *coefficient of determination*:

$$r^2 = \frac{s_{Y_c - \bar{Y}}^2}{s_Y^2} = \frac{\text{explained variance}}{\text{total variance}}$$

The coefficient of determination is defined in the above equation as the proportion of the total variance in the dependent variable which is explained by the independent variable. The coefficient of determination is preferred to the coefficient of correlation for most applications in business and economics because it is a more clear-cut way of stating the proportion of the variance in Y which is associated with X . The coefficient of correlation may suggest a higher degree of correlation than really exists. Thus, if 50 percent of the variance in Y is explained by X (and the other 50 percent is not explained), $r^2 = 0.50$, but $r = \sqrt{0.50} = 0.71$.

The coefficient of determination may also be expressed as 1 minus the proportion of total variance which is *not* explained. That is,

$$r^2 = 1 - \frac{S_{YX}^2}{S_Y^2} = 1 - \frac{\text{unexplained variance}}{\text{total variance}}$$

This formula is more convenient for computation than the first one, since the unexplained variance is the square of the standard error of estimate (S_{YX}), which we have already computed in regression analysis.

Thus, in the production rating case:

$$\text{Unexplained variance is } S_{YX}^2 = (5.13)^2 = 26.3 \quad (\text{page 562})$$

$$\text{Total variance is } s_Y^2 = \frac{\Sigma y^2}{n - 1} = \frac{7,050}{19} = 371 \quad (\text{Table 22-2})$$

$$r^2 = 1 - \frac{26.3}{371} = 0.929$$

That is, 92.9 percent of the variance in production ratings is explained, or accounted for, by the variance in test scores; only 7.1 percent of the variance is not so explained. The correlation coefficient is

$$r = \sqrt{0.929} = 0.964$$

The correlation coefficient for a sample may also be defined by the following formula:

$$r_s = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

The term Σxy measures the degree to which x and y vary with each other, and the terms Σx^2 and Σy^2 measure the individual variation in X and Y , respectively. The correlation coefficient is thus a measure of the covariation of X and Y relative to the variation of X and Y themselves.

In certain preliminary studies, and particularly in the application of psychology to business problems, a relative measure of degree of relationship between X and Y may be all that is needed. For example, an industrial psychologist may be interested in finding which factors are related to the morale of a group of employees. He may not be interested in explicitly predicting employee morale from the other factors. Thus, he may not wish to use regression analysis, but may still use the correlation coefficient to measure the degree of the relationship between morale and each of the other factors.

Note that the above formula also provides a short-cut method for

calculating the coefficient of determination and the coefficient of correlation.

In the production rating case (Table 22-2):

$$r_s^2 = \frac{(6,974)^2}{7,395 \times 7,050} = 0.933$$

This sample value, however, is biased as an estimate of the true population value of r^2 . The best estimate of the latter is, in this example,

$$r^2 = 1 - (1 - r_s^2) \left(\frac{n-1}{n-2} \right)$$

$$r^2 = 1 - (1 - 0.933) \left(\frac{19}{18} \right) = 0.929$$

This is the same result as in the formula⁷

$$r^2 = 1 - S_{YX}^2 / S_Y^2.$$

Graphic Analysis

The coefficient of determination may also be estimated graphically by use of the preceding formula. The method is illustrated in Chart 22-11. This chart shows the effect of weight on handling time for 22 pieces of metal in a time study of an operation at a John Deere plant. The purpose of this study was to determine the best sizes of metal stock to use in feeding the bump gauge of a punch press.

The procedure is as follows: First, plot a large-scale scatter diagram and fit a freehand regression line, as described earlier in the chapter.

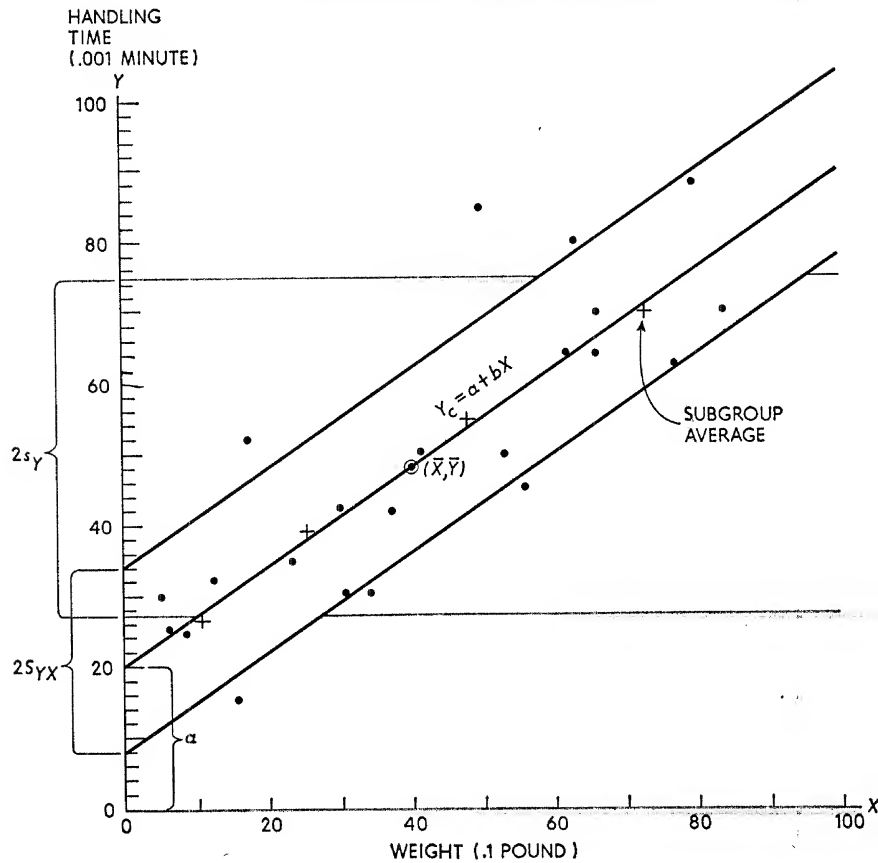
Second, draw two lines parallel to the regression line so that one sixth of the dots fall above and one sixth below this band. Thus, if there are 22 points as in Chart 22-11, the line may be drawn between the third and fourth dots from the top and bottom, measured toward the regression line. This may be done with a transparent ruler or parallel rules set along the regression line. In the case of a curved line, trace the curve and the Y axis on a transparent sheet and move this sheet up and down along the Y axis until one sixth of the dots are excluded on either side.

Now measure the vertical width of this band on the Y axis. This value is roughly twice the standard error of estimate, $2S_{YX}$, since a range of S_{YX} above and below the regression line includes about two thirds of

⁷ In this formula, we adjusted for sample bias by using $n-2$ and $n-1$, instead of n , in computing S_{YX} and S_Y , respectively, to compensate for the loss of degrees of freedom in measuring deviations from the regression line and \bar{Y} .

Chart 22-11

WEIGHT AND HANDLING TIME OF 22 PIECES OF METAL
FED TO BUMP GAUGE OF NO. 13 PUNCH PRESS



Note: The bands on the chart are drawn horizontally and parallel to the regression line so as to exclude one sixth of the points on either side.
Source: John Deere and Company.

the items in a normal distribution. In Chart 22-11, $2S_{YX}$ is about 26, so S_{YX} is 13.

If gaps occur in the data near either of the points marked, the band may be drawn to exclude a fifth or some other fraction of the dots on either side, provided the same number of points falls outside the horizontal band in step 3, below. Since r^2 depends on the *ratio* of the two scatters, the proportion of points excluded might vary considerably without impairing the accuracy of this ratio.

Third, set the ruler on the scatter diagram *horizontally* and mark two straight lines separating off the top sixth of the items and the bottom

sixth. Measure this spread, too, against the vertical scale of the chart. This is roughly $2s_Y$, or twice the standard deviation of the dependent variable, since a range of s_Y above and below the mean of the Y values includes about two thirds of the items in a normal distribution. Here, $2s_Y$ is about $47\frac{1}{2}$, so $s_Y = 23\frac{3}{4}$.

Finally, substitute these values in the above formula. In this example,

$$\begin{aligned} r^2 &= 1 - \frac{S_{YX}^2}{s_Y^2} \\ &= 1 - \frac{(13)^2}{(23.75)^2} \\ &= 0.70 \end{aligned}$$

This measure of correlation is useful as a quick estimate of r^2 or as a check on the computed value. It is relatively accurate when r^2 is high. The chart also provides a visual picture of the degree of correlation: the smaller the *ratio* of the sloping band to the horizontal one, the higher the correlation.

Sampling Error of the Correlation Coefficient

We will not take up the standard error of the correlation coefficient directly, since this concept involves difficulties that are disproportionate to its rather limited usefulness in business.⁸

The sampling variability of correlation coefficients may be illustrated graphically, however, in Chart 22-12. This chart shows the *minimum* value of the true correlation coefficient for any sample value of r , at the 95 percent confidence level.

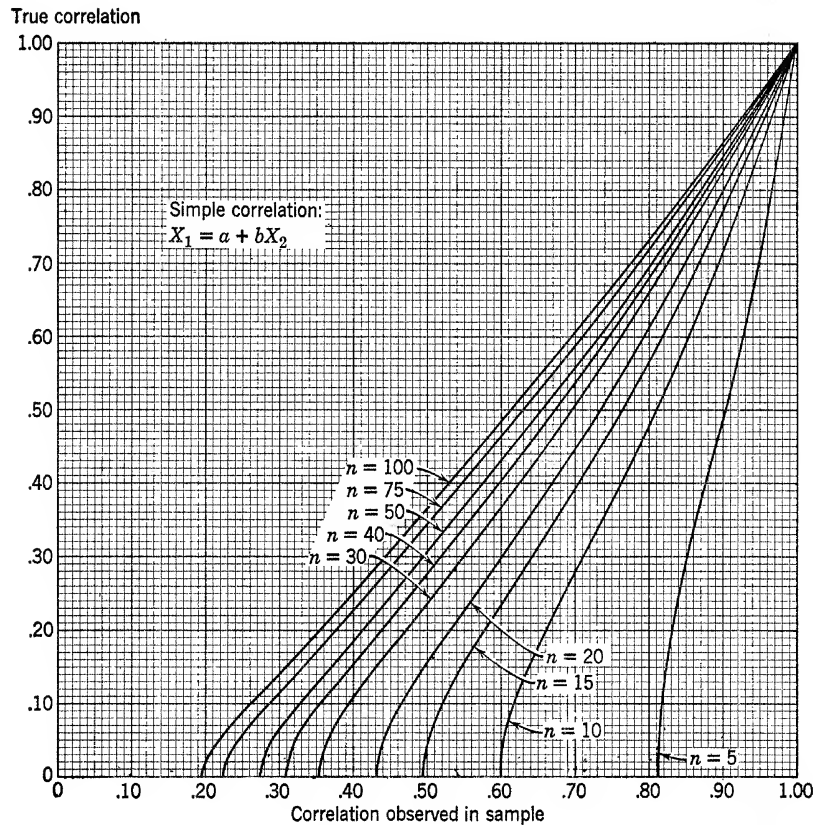
For example, in the production rating case, the coefficient of correlation for the example of 20 workers is $\sqrt{.929}$, or .964. With this value on the X axis, use the $n = 20$ curve to find .93 on the Y axis. We can say, therefore, that the true correlation for the population is *at least* .93, with a 95 percent chance of being correct.

If the sample r were .60, however, with $n = 10$, we could only say that the true value is at least zero, with the same degree of confidence. That is, even if there is *no* correlation in the population itself, 5 percent of all possible samples of size 10 would still yield a correlation coeffi-

⁸ The standard error of the correlation coefficient can be estimated as $s_r = (1 - r^2) \div \sqrt{n - 1}$. This formula is only applicable to large samples, and even then the distribution of the sample r 's is quite skewed when the true value of r is far from zero. The value r , however, can be transformed into a quantity called Fisher's z , whose sampling distribution is nearly normal. For a treatment of confidence intervals and tests of hypotheses using z , see W. A. Spurr, L. S. Kellogg, and J. Smith, *Business and Economic Statistics* (Homewood, Illinois: Richard D. Irwin, 1954), pp. 492-93, and Appendix I.

Chart 22-12

MINIMUM CORRELATION IN POPULATION, FOR VARYING
OBSERVED CORRELATIONS AND SIZE OF SAMPLE



Under conditions of random sampling, one sample out of 20, on the average, will show a correlation coefficient with a \pm value as high as that "observed in sample," when drawn from a population with the stated true correlation. Reprinted with permission from M. Ezekiel, and K. A. Fox, *Methods of Correlation and Regression Analysis* (3d ed.; New York: John Wiley, 1959), p. 294.

cient of $\pm .60$ or higher. This chart demonstrates the danger of making inferences about the degree of correlation when r or n is small.

EXAMPLES OF REGRESSION ANALYSIS

In this section we shall give a few brief examples of the use of regression analysis in business decision-making.

Regression for Prediction

Work scheduling in a mail-order house is dependent upon knowing how many orders will arrive for processing each day.⁹ This information

⁹This illustration is based on the article "Estimating Daily Order Receipts from Weight of Mail," by C. M. Smalley, in *American Statistician* (February 1954).

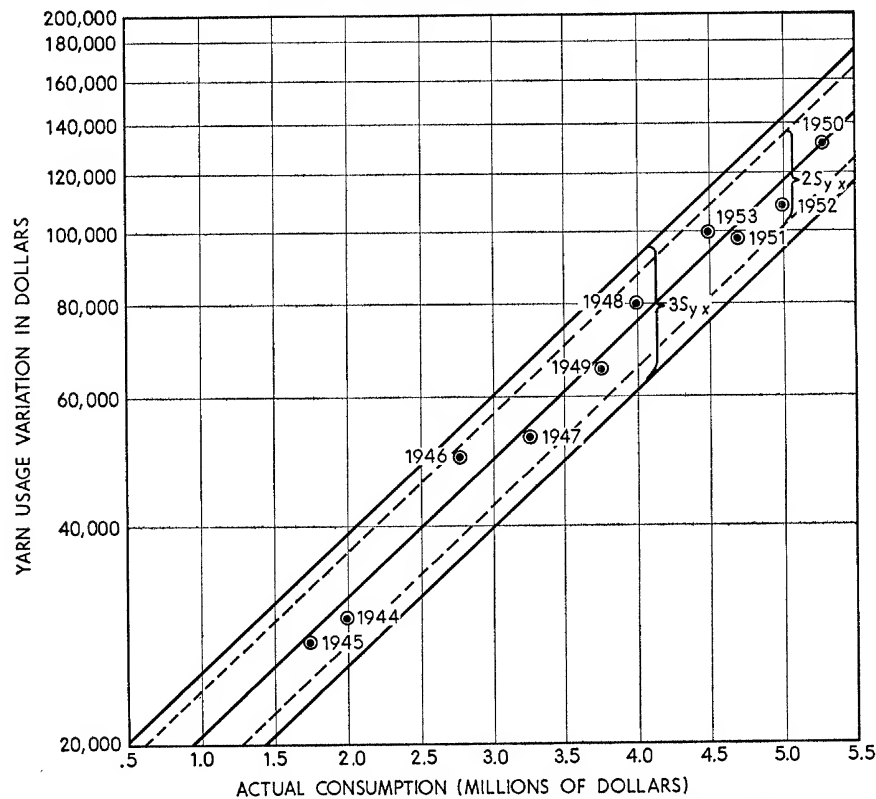
is needed early in the day, before the incoming mail is sorted, opened, and classified. Mail-order houses have solved this problem by using the weight of the mail as a means of estimating the number of orders. The mail is quickly weighed. Using the past linear relationship between weight of mail and number of orders, the latter is easily estimated each morning. But the relationship between the weight and the number of orders varies for different days of the week. For example, Monday generally has fewer orders per pound of mail than Tuesday. Hence, a different regression line is used for each day of the week. Mail-order houses have found this a reliable and efficient means of estimating daily orders.

Regression for Control

In cost accounting, management reports include the planned or "standard" cost for a given activity, plus a "variance" or deviation from

Chart 22-13

RELATIONSHIP BETWEEN YARN USAGE VARIATION AND
ACTUAL CONSUMPTION OF YARN



SOURCE: A. W. Patrick, "A Proposal for Determining the Significance of Variations from Standard," *The Accounting Review* (October 1957), p. 590.

the standard. (This usage of "variance" is entirely different from its use in statistics.) If the variance is large, management will investigate to determine the cause. If the variance is small, it can be attributed to minor factors, so no investigation is necessary. This leaves unanswered the question of how large a variance must be before an investigation is undertaken.

In order to answer this question, we first find the past relationship between the planned cost and the variance for a given activity. A certain variance can be related to the regression line to determine if it is "out of line" with other points. An observation that is more than two or three standard errors of estimate above or below the line is likely to need investigation.

An example is shown in Chart 22-13.¹⁰ The actual consumption of yarn is plotted against the accounting variance in yarn usage. The regression line has been calculated on the basis of past data and bands are drawn at a distance of two and three S_{YX} . Points falling outside these lines call for careful investigation.

In this example regression analysis is used as a means of management control over costs.

CAUTIONS IN THE USE OF CORRELATION AND REGRESSION ANALYSIS

Before concluding this chapter, it is well to point out some pitfalls that may trap the unsuspecting in their use of regression and correlation.

Curvilinear Relationships

Throughout this chapter, we have assumed that the data fit a straight line. If the points are plotted on a scatter diagram, this assumption can be verified easily. When there are many points, and especially in using a computer program, the unwary may skip the step of plotting the data as a check on the linearity assumption. Beware of this pitfall, since it may lead to very poor predictions. If there are a large number of points, at least a sample of them should be plotted. There are also mathematical formulas for checking the assumption of linearity.¹¹ Methods of handling curvilinear relations are discussed in Chapter 24.

The assumptions that the points have a uniform, random scatter about the regression line should also be tested before making any

¹⁰ The example is based on the article by A. W. Patrick, "A Proposal for Determining the Significance of Variations from Standard," *The Accounting Review* (October 1957).

¹¹ See, for example, W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis* (New York: McGraw-Hill, 1957), pp. 197-98.

prediction or inference. These assumptions can generally be verified by a visual check of the plotted data.

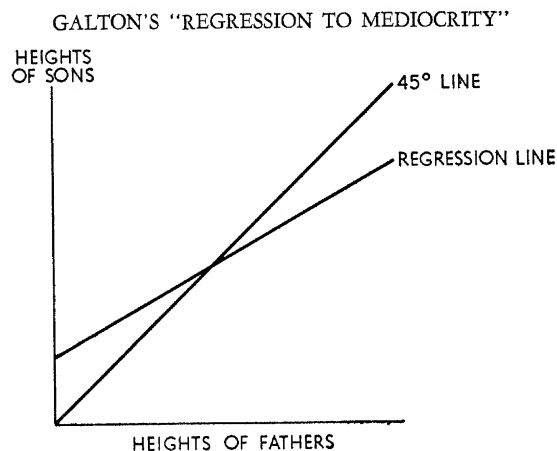
Correlation and Causation

The fact that two variables are correlated does not imply in any way that either is a cause of the other. As noted in Chapter 1, if X and Y are correlated, it may be that (1) X causes Y , (2) Y causes X , (3) X and Y interact on each other, (4) both are influenced by Z , or (5) the correlation is due to chance. To cite a *reductio ad absurdum*, church attendance and beer consumption correlate over the years, but this does not mean that attending church makes one thirsty or that drinking beer incites piety; they have both simply increased with population growth. Other examples are cited in Chapter 1. A whole branch of statistics is concerned with the design and analysis of experiments to control extraneous factors and determine underlying causal relationships.

Regression Fallacy

The regression fallacy is pervasive and insidious. It was noted by Sir Francis Galton, when he plotted the heights of fathers against the heights of their sons, that the line of average relationship had a less inclined slope than the expected 45° line (Chart 22-14). That is, very *tall* fathers had sons *shorter* than they, whereas *short* fathers had sons *taller* than they, on the average. Galton termed this phenomenon a "regression to mediocrity" in height from one generation to the next, thus giving rise to the inapt statistical term "regression." The same phenomenon has been noted in company profits, examination scores,

Chart 22-14



results of advertising campaigns, and almost any variable that one attempts to correlate with itself at a previous time. A whole book has been written deploring "The Triumph of Mediocrity in Business," based upon such an analysis of company profits and sales.

The fallacy in this reasoning arises from the fact that a series of values usually fluctuates around its average or trend level from time to time. At any particular time, some of the highest values reflect nonrecurring factors (e.g., a company's windfall profits), which are usually followed by more normal values in the succeeding period. In Galton's example, an unusually tall member of the male line of descent is likely to have a son of more normal height. By the same token, a particularly brilliant father will typically have sons of more moderate ability; the sons should not chide themselves for their supposed "failure." Just as high values may be abnormal, low values may reflect an unusual combination of depressing causes, and so tend to be followed by more middling values. Hence, any series that fluctuates is apt to show this spurious convergence toward the mediocre. The proper way to determine whether such convergence exists is not to use regression analysis but to compare the dispersion of the data in the two periods.¹²

SUMMARY

Simple correlation and regression analysis is concerned with the study of two variables and how they change together from observation to observation. The variables should be carefully chosen in such a way that there is a meaningful interpretation of the relationship between them.

In most such studies, interest is concentrated on estimating one variable from the other. The one to be estimated is called the *dependent* variable Y , and the other is called the *independent* variable X . These are plotted on a scatter diagram, which shows whether the relationship is close or not, whether it is positive or negative, and whether it is linear or curvilinear.

The basic measures of relationship are the *regression line* or curve, which describes the average relationship between X and Y ; the *standard error of estimate*, which is the standard deviation of the residuals ($Y - Y_c$) around this line; and the *coefficient of correlation*, a relative measure of relationship which varies from 0 to ± 1 .

Regression analysis is used in business and economics principally for the purposes of *prediction* and *control*. Thus, in correlating the earnings per share (X) with price per share (Y) for a number of stocks, we can

¹² See W. A. Wallis and H. V. Roberts, *Statistics, A New Approach* (New York: The Free Press, 1956), pp. 258-63, for a further discussion of this fallacy.

predict the price of a stock from the regression line, based on estimated future earnings, or we can use the standard error of estimate to construct a confidence interval around this line and consider the stock unduly high or low in price if it is outside these control limits.

Regression lines or curves can be fitted either graphically or mathematically. In graphic analysis, arrays are constructed by grouping observations for which values of X are approximately equal; a point of means for each array is estimated and indicated by a small cross or circle; and a smooth curve is drawn to fit the points of means. Such a curve should be relatively inflexible. If the regression is linear, the line is drawn through (\bar{X}, \bar{Y}) , the point of means of all observations.

The regression of Y on X is said to be *linear* or *curvilinear*, depending on the shape of the curve determined by the means of arrays of Y values for various values of X . When the regression is linear, the two constants of the regression line are its Y intercept a and its slope b , the *regression coefficient*.

The *method of least squares* is a means of computing the constants of the regression line so as to minimize the sum of squares of residuals from the line. Thus, in fitting a straight line, $\Sigma(Y - Y_c)^2$ is less than for any other straight line. A straight line fitted by least squares also goes through the overall means of the data and reduces the sum of the plus and minus deviations to zero: $\Sigma(Y - Y_c) = 0$. The computations can be simplified by using the deviations of the variables from their means (i.e., using x and y instead of X and Y).

The *standard error of estimate* measures the average error of the regression line in providing estimates of Y from given values of X . It may be computed as the standard deviation of the residuals $(Y - Y_c)$ around the regression line or by means of a short-cut formula.

When the data used for regression analysis can be considered as a random sample from a population, we can make statistical inferences based upon the sample data. The assumptions in linear regression analysis are (1) linear relationship between X and Y in the population; (2) uniform scatter about the regression line; (3) the independence of the deviations about the regression line; and (4) a roughly normal distribution of points about the regression line. When these assumptions are satisfied, the sample values a and b are "best" estimates of the population values A and B .

We should also distinguish between the *correlation* model and the *regression* model. In the correlation model, both X and Y are assumed to be normally distributed and all correlation and regression statistics are valid estimators. In the regression model, the Y values are normally distributed, but the X values may be arbitrarily limited, as in a con-

trolled experiment. In this case, regression results are valid only for these same X values, and the correlation coefficient is not generally valid.

We can apply *tests of significance* and *confidence intervals* to regression results from random samples in order to make statistical inferences about the parent population. Thus, we can determine whether there is any significant relationship between X and Y by testing the null hypothesis that the population regression coefficient B is zero. If the sample value b , divided by its standard error, is sufficiently large, according to a table of the normal or t distribution, the relationship is deemed to be significant.

Using the standard errors of the regression coefficient (s_b) and the regression line (s_{Y_e}), we can compute confidence intervals for the regression coefficient and the regression line, respectively. By further combining the standard error of the regression line with the standard error of estimate, we obtain the *standard error of forecast*, which provides confidence limits within which any *new observation* may be expected to fall. The confidence bands for both the regression line and an individual forecast are narrowest at \bar{X} ; they widen out in either direction. This indicates the danger of estimating Y for values of X that are far from their mean. The forecast error is valuable both in predicting Y and in providing a control chart for Y .

The *coefficient of correlation* is a relative measure of relationship. Its square, the *coefficient of determination*, is the ratio of explained variance to total variance, or 1 minus the ratio of unexplained to total variance.

Total variance is the standard deviation (squared) of the Y values around their mean ($Y - \bar{Y}$). *Explained variance* is the standard deviation (squared) of the Y_e values around the mean ($Y_e - \bar{Y}$), since this part of the variation in Y can be explained by corresponding changes in X . *Unexplained variance* is the standard deviation (squared) of Y values around the regression line ($Y - Y_e$)—the variation in Y not explained by X . This is the standard error of estimate, squared. The coefficient of determination is a more direct and unequivocal measure of the proportion of variance in Y explained by X than is the higher-valued coefficient of correlation.

The coefficient of determination may be estimated graphically from the ratio of the vertical widths of two bands drawn horizontally and parallel to the regression line—each including the central two thirds of the dots. It may also be computed directly by a short-cut formula. Confidence limits for r are shown in Chart 22-12. The chart illustrates the dangers of making inferences when r or n is small.

In conclusion, the regression coefficient b , the standard error of

estimate S_{YX} , and the coefficient of determination r^2 each measure a different aspect of a given relationship. In the production rating example, the regression coefficient tells us the average *amount* of change in production for a given change in the test score; the standard error of estimate tells us how *accurate* is our estimate of production; and the coefficient of determination tells us what *proportion* of variance in production ratings is accounted for by the test scores. For many problems of control and prediction, the first two measures will suffice. The coefficient of determination is needed only if the problem calls for a measure of *proportionate* importance.

Three pitfalls in the use of regression analysis should be noted: (1) the data should always be plotted or otherwise checked to avoid using linear regression analysis on curvilinear data; (2) correlation between two variables does not, of itself, imply that there is any causal relationship between the variables; and (3) the regression fallacy occurs when a variable is plotted against itself in a previous time period. Chance variation, and not any "regression toward mediocrity" causes the regression line to incline below the 45° line.

PROBLEMS

1. Distinguish between:
 - a) Regression and trend analysis.
 - b) Linear and curvilinear regression.
 - c) The standard error of estimate and the standard deviation of the dependent variable.
 - d) The use of regression analysis for prediction and for control.
 - e) The coefficient of regression and the coefficient of correlation.
 2. Explain:
 - a) The method of least squares, as applied to regression analysis.
 - b) How to test whether there is any significant relationship between two variables.
 - c) How to obtain a 99 percent confidence interval for the regression coefficient in a large sample.
 - d) How the standard error of forecast is derived from the standard error of estimate and the standard error of the regression line.
 - e) The coefficient of determination in terms of explained variance, unexplained variance, and total variance.
 3. Answer the following questions by inspection of Chart 22-11.
 - a) Is the relationship between weight and handling time simple or multiple, linear or curvilinear, positive or negative, significant or negligible?
 - b) Give the approximate regression equation. Explain the meaning of the a and b values in estimating handling time from weight.
-

- c) Give the estimated handling time (Y_c) for pieces weighing 80 tenths of a pound. What is the unexplained variation ($Y - Y_c$) for the piece that weighed this amount but actually required 88 thousandths of a minute to handle?
- d) Considering the sampling error of the regression line as well as the standard error of estimate, for what weight could you forecast handling time most accurately?
4. Assume that we conduct an experiment with eight fields planted to corn: four fields having no nitrogen fertilizer and four fields having 80 pounds of nitrogen fertilizer. The resulting corn yields are shown in the table, in bushels per acre.

Field	Nitrogen (Pounds)	Corn Yield Bushels/Acre
1.....	0	12
2.....	0	36
3.....	0	6
4.....	0	18
5.....	80	128
6.....	80	112
7.....	80	112
8.....	80	72
Totals.....	320	496

Note: This sample is too small to provide really valid inferences, but it serves to illustrate the methods involved with a minimum of computations.

- a) Plot the data as a scatter diagram on an arithmetic chart, and draw a regression line by the graphic method, using group averages as guides.
- b) Compute a linear regression equation by least squares. How does this compare with the graphic line when plotted on the chart? Explain the meaning of the regression equation in terms of fertilizer and corn yields.
- c) Compute the standard error of estimate. Interpret this value in terms of predicting corn yields.
- d) Predict corn yield for a field treated with 60 pounds of fertilizer, and give the 95 percent confidence limits for this prediction. (Assume a linear relationship and ignore sampling errors in the regression line itself.)
- e) Compute the estimated coefficient of determination as 1 minus the unexplained variance over the total variance. What does this figure tell you about the relationship of nitrogen fertilizer and corn yields in general?
5. Refer to the data described in Problem 4.
- a) Is there any significant relationship between nitrogen fertilizer and corn yields? That is, test the null hypothesis $B = 0$ against the alternative hypothesis $B > 0$ at a critical probability of, say, 5 percent.
- b) Give the 95 percent confidence interval for the regression coefficient.
- c) How is your interpretation of the results in *a* and *b* affected by the fact that the basic data represent a controlled experiment rather than a survey in which both X and Y are normally distributed? (Ignore the small size of the sample.)

6. In the same corn-yield experiment (Problems 4 and 5 above):
- Compute the standard error of the regression line and its 95 percent confidence limits for fertilizer applications of 0, 40, and 80 pounds, respectively.
 - Compute the standard error of forecast and the 95 percent confidence limits for individual forecasts of corn yield, assuming fertilizer applications of 0, 40, and 80 pounds, respectively.
 - How is your interpretation of the results in *a* and *b* affected by the fact that the basic data represent a controlled experiment rather than a survey in which both *X* and *Y* are normally distributed? (Ignore the small size of the sample.)
7. *a*) Estimate the coefficient of determination for test scores and production ratings in Chart 22-6 by the graphic method. How does this result compare with the computed value of $r^2 = 0.93$?
- b*) If the sample value of r had been 0.60 in this example, with $n = 20$, what is the minimum value of the true correlation coefficient of the population at the 95 percent confidence level (Chart 22-12)?
- c*) If the true correlation coefficient were zero, what sample value would be exceeded by 5 percent of all random samples of size 20?
8. Refer to Table 23-3. Consider the simple regression between the area of a lot (*X*) and its price (*Y*).
- Verify that the least squares regression equation is $Y_e = 1.453 + 0.2194 X$. (Refer to Table 23-5.)
 - Is the relationship between area and price statistically significant?
 - Calculate the correlation coefficient between area and price.
 - A given lot has 18,000 square feet. Estimate the price at which it sold. Give a 95 percent confidence interval about this estimate.
9. Refer to Tables 23-3 and 23-5.
- Estimate the simple regression line between the elevation of a lot and its price.
 - Calculate the standard error of estimate.
 - Is the relationship between elevation and price significant?
 - Calculate the correlation coefficient between elevation and price.
10. An analyst for a certain company was studying the relationship between travel expenses in dollars (*Y*) for 102 sales trips and the duration in days (*X*) of these trips. He has plotted the data, and the relationship is approximately linear. The data are summarized in the table.

	X	Y	X ²	XY	Y ²
Totals	510.0	7140.0	4150.0	54,900.0	740,200.0
Means	5.0	70.0			
Adjustments			-2550.0	-35,700.0	-499,800.0
Adjusted totals			1660.0	19,200.0	240,400.0
which is			Σx^2	Σxy	Σy^2

- a) Estimate the regression equation from the above data.
 - b) What is the practical significance of the value of a (the intercept) in this equation?
 - c) A given trip is to take seven days. How much money should a salesman allow so that there is only one chance in ten that he will run short?
11. The Scuffo Shoe Company operated a chain of retail shoe stores. As a means of measuring the efficiency of the various stores, a study was made of the relationship between the number of employees (X) and the average monthly sales volume (Y) for all the stores over the past year. When the data were plotted, the relationship was approximately linear, with the points having a uniform scatter about the line. The data can be summarized as follows: X = the number of employees in each store; Y = the average monthly sales during 1966 for each store in thousands of dollars; $n = 100$ = the number of stores in the Scuffo chain; $\Sigma X = 600$; $\Sigma Y = 1,600$; $\Sigma X^2 = 5,200$; $\Sigma Y^2 = 37,700$; $\Sigma XY = 13,600$.
- a) Find the line of average relationship (i.e., the regression line). Give a verbal meaning of this equation.
 - b) Calculate the coefficient of correlation.
 - c) Store No. 86 employs 10 persons and has monthly sales of \$20,000. Is the performance of this store "out of line" with the performance of the other stores? How do you know?
12. As the Alma Mater University Alumni secretary in your city, you are responsible for making reservations for the semimonthly alumni luncheons. Before each meeting you send out letters with return postcards. Each alumnus is asked to return this card if he plans to attend. You find that only a portion of the cards are returned by the time it is necessary to make the reservation, and you are forced to guess about the actual number of lunches that will be necessary.
- You have analyzed the data over the past two years (48 luncheons) and have found that there is approximately a linear relationship between the number of reservations received (by four days before the luncheon) and the actual number present at the luncheon. Therefore, you fit a regression line to the data and find: $Y_e = 20 + 1.50 X$, where Y_e is the estimate of the actual attendance and X is the number of reservations received by four days before the luncheon. You also have $S_{YX} = 5.0$; $n = 48$; $\bar{X} = 20.0$; $\Sigma x^2 = 4,700$; $\bar{Y} = 50.0$; $\Sigma y^2 = 10,575$; $\Sigma xy = 7,050$.
- a) Explain the meaning of the regression equation above.
 - b) Suppose 38 reservations are received for a given luncheon. Calculate a forecast interval at the 95 percent confidence level. (Assume that the deviations about the regression line are normally distributed.)
13. Refer to the data in Table 14-5. Calculate the correlation coefficient between current inventory and annual inventory on an item basis. What is the minimum correlation in the whole population at the 95 percent level? (Use Chart 22-12.)

14. A certain mail-order firm used the weight of the incoming mail to estimate the number of orders that would need to be processed. Over a 25-day period the following data were collected:

<i>Day No.</i>	<i>Weight of Mail (Hundreds of Pounds)</i>	<i>Thousands of Orders</i>	<i>Day No.</i>	<i>Weight of Mail (Hundreds of Pounds)</i>	<i>Thousands of Orders</i>
1	1.8	6.4	14	4.1	13.8
2	2.0	8.0	15	4.2	12.8
3	2.0	7.2	16	4.2	16.5
4	2.1	7.5	17	4.2	17.1
5	2.3	6.9	18	4.3	15.4
6	2.6	10.9	19	4.6	16.2
7	2.6	10.3	20	5.0	15.8
8	2.8	9.5	21	5.4	19.0
9	3.1	9.7	22	5.8	19.4
10	3.2	10.6	23	6.0	19.1
11	3.2	12.5	24	6.4	18.5
12	4.0	12.9	25	6.5	20.0
13	4.1	14.0			

- Calculate the linear regression equation relating the number of orders to the weight of the mail.
 - What is the sampling error associated with the estimated slope b ? Are you sure that the true value B is greater than 2.5?
 - Estimate the number of orders for a mail delivery that weighs 500 pounds.
 - Assuming that the points are approximately normally distributed about the regression line, place 95 percent forecast limits on the estimate calculated in c above.
15. Wheat yields in Nebraska have a total variance of 25 bushels per acre over many years, of which a variance of 16 bushels can be explained by variations in seasonal rainfall. This year's yield is estimated at 26 bushels an acre (near the long-term average) based on the season's rainfall of 18 inches.

Within what range would you predict the yield to be this season, on a given farm, with about 95 chances out of 100 of being correct?

SELECTED READINGS

Selected readings for this chapter are included in the list which appears on page 657.

23. MULTIPLE CORRELATION AND REGRESSION

MULTIPLE CORRELATION and regression analysis enables us to measure the joint effect of any number of independent variables upon a dependent variable. The multiple regression equation describes the average relationship between these variables, and this relationship is used to predict or control the dependent variable. The standard error of estimate is essentially the standard deviation of this variable from its computed values. And, finally, the coefficient of multiple determination measures the proportion of the variance in the dependent variable explained by the other factors. The concepts and techniques in this chapter, therefore, are just extensions of those in simple correlation. However, by measuring the simultaneous influence of several factors, we have a more powerful and realistic tool of analysis than in considering only one independent variable, and the use of computer programs facilitates the calculations.

To illustrate the use of several variables, consider the problem of predicting new automobile sales for the coming year. There are many factors that affect sales, each one explaining a part of the total. Plausible factors include the number of existing motor vehicles registered at the end of the current year; the average age of existing automobiles; the total population 16 years of age or older; the level of disposable personal income per capita; and the expected retail prices for new automobiles relative to the general price level for consumer goods and services. Here, common sense (*and* economic theory) should indicate whether each of these variables has a positive or a negative effect upon the sales of new automobiles. It would appear that at least five independent variables would be necessary to explain or forecast variations in the sales of new automobiles.

Multiple regression is often used in connection with forecasting. Such a forecast may be as broad as the general economic outlook for the nation as a whole, or it may be limited to the estimation of the price of a single stock. For example, the Value Line Investment Survey correlates the price of a stock in past years with its earnings per share and dividends (all in logarithms) to determine the estimated future value of the stock. Recommendations for stock purchase are based in part on this "value line" obtained by multiple regression analysis.

This chapter is concerned only with rectilinear multiple regression analysis, in which each independent variable is assumed to have a linear relationship with the dependent variable. Curvilinear relations are discussed in the following chapter.

MULTIPLE REGRESSION

The multiple regression equation represents the simultaneous influence of a set of independent variables upon the dependent variable. The linear equation can be written as

$$Y_c = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots$$

where Y_c is the computed or estimated value of the dependent variable Y and X_1, X_2, X_3, \dots are the independent variables. The equation is said to be linear (or rectilinear) since there are no terms such as X_1^2 or X_1X_2 present. The term a is simply the value of Y_c when all the X 's are zero. The terms b_1, b_2, b_3, \dots are the *net regression coefficients*. Each measures the change in Y per unit change in that particular independent variable. However, since we are measuring the simultaneous influence of all variables on Y , the net effect of X_1 (or any other X) must be measured apart from any correlated influence of other variables. This is usually expressed by adding the qualifying statement: "All other variables held constant" or "adjusting for the effect of the other variables." We would say, therefore, that b_1 measures the change in Y per unit change in X_1 , *holding the other independent variables constant*.

To illustrate, suppose we wish to predict job performance (Y) of applicants for a given job based on the score of a placement test (X_1) and the interviewer's rating (X_2). The scales are arbitrary. We test a random sample of 18 new employees and later measure their job performance.

In Table 23-1 it can be seen that each successive pair of observations provides a set of values of Y for which X_1 and X_2 are constant. Means of these sets of Y values are presented in Table 23-2. When X_1 increases by 10, the mean of Y increases by 4 (four tenths as much as

Table 23-1

RELATION OF TEST SCORES AND INTERVIEWER'S RATINGS
TO JOB PERFORMANCE (18 EMPLOYEES)

Employee Number	Job Performance Y	Test Score X_1	Interviewer's Rating X_2
1	5	10	5
2	13	10	5
3	9	20	5
4	17	20	5
5	13	30	5
6	21	30	5
7	14	10	20
8	22	10	20
9	18	20	20
10	26	20	20
11	22	30	20
12	30	30	20
13	20	10	30
14	28	10	30
15	24	20	30
16	32	20	30
17	28	30	30
18	36	30	30
Total	378	360	330
Mean	21	20	18.33

X_1), and as X_2 increases by 15 or 10, the mean of Y increases by 9 or 6, respectively (six tenths of the change in X_2). Accordingly, the net regression coefficients are $b_1 = 0.4$ and $b_2 = 0.6$. In order to determine the intercept value a , note that the regression plane must go through the overall means of the data. Hence,

$$\bar{Y} = a + b_1\bar{X}_1 + b_2\bar{X}_2$$

or

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 21 - (0.4)20 - (0.6)(18.33) = 2$$

Hence, the regression equation is

$$\begin{aligned} Y_c &= a + b_1X_1 + b_2X_2 \\ &= 2 + 0.4X_1 + 0.6X_2 \end{aligned}$$

Table 23-2
MEANS OF ARRAYS OF THE DEPENDENT VARIABLE Y

	$X_2 = 5$	$X_2 = 20$	$X_2 = 30$
$X_1 = 10$	9	18	24
$X_1 = 20$	13	22	28
$X_1 = 30$	17	26	32

SOURCE: Table 23-1.

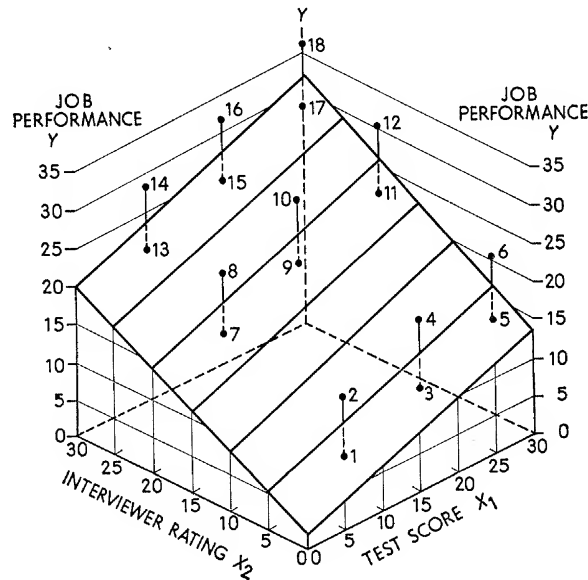
The net regression coefficient b_1 shows the average effect of a one-unit increase in X_1 (test score) on Y (job performance), *holding* X_2 *constant*. That is, b_1 indicates how the test score predicts job performance for men rated alike by the interviewer. The net regression coefficient thus differs from the *gross* regression coefficient b in simple correlation between test scores and job performance in that b shows the *combined* effect of test score and the intercorrelated effect of interviewer's rating in predicting job performance.

The regression equation above is the equation of a plane in three-dimensional space, as shown in Chart 23-1. The observed points scatter above and below the plane. For *linear* multiple regression, we assume

Chart 23-1

MULTIPLE REGRESSION PLANE

$$Y_c = 2 + 0.4X_1 + 0.6X_2$$



that such a plane is a good fit to the data. If not, some curvilinear surface may be more appropriate (see Chapter 24).

ESTIMATION OF MULTIPLE REGRESSION COEFFICIENTS

The multiple regression coefficients may be estimated either by graphic or least squares method. Today, electronic computers provide a variety of fast and accurate programs for least squares analysis. However, graphic techniques are useful (1) in understanding the basic concepts of multiple regression, (2) to check the assumptions underlying this analysis (e.g., linearity and homoscedasticity), (3) to obtain quick results when no computer is available, and (4) to determine curvilinear relationships (Chapter 24) when the appropriate equation form is unknown. For these reasons we shall briefly present the graphic method before discussing the least squares technique.

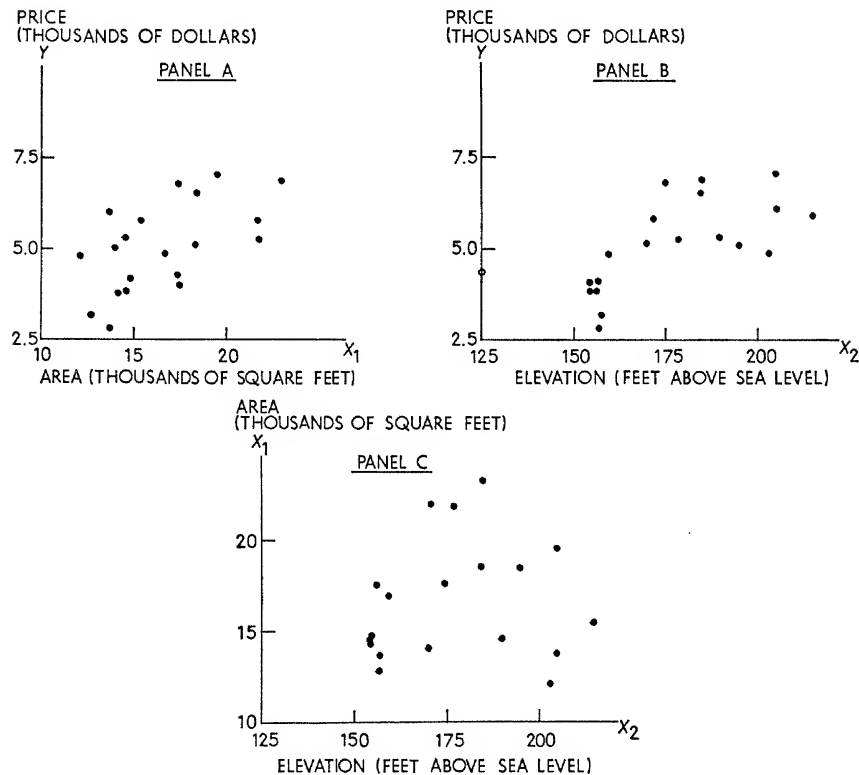
Table 23-3

AREA, ELEVATION, AND PRICE FOR 20 RESIDENTIAL LOTS

Lot No.	X_1 Area, Thousands of Square Feet	X_2 Elevation, Feet above Sea Level	Y Price, Thousands of Dollars
1	14.7	155	4.1
2	14.2	155	3.9
3	12.7	158	3.2
4	13.8	158	2.9
5	14.4	155	3.9
6	17.4	157	4.1
7	21.8	172	5.8
8	14.0	170	5.1
9	17.5	175	6.8
10	23.0	185	6.8
11	18.3	185	6.5
12	19.4	205	7.0
13	15.2	215	5.8
14	18.3	195	5.1
15	21.7	178	5.3
16	16.7	160	4.9
17	13.6	205	6.0
18	14.5	190	5.3
19	12.1	203	4.8
20	17.4	125	4.3
Total	330.7	3501.	101.6
Mean	16.535	175.05	5.08

Chart 23-2

RELATION BETWEEN AREA, ELEVATION, AND PRICE OF 20 LOTS
SCATTER DIAGRAMS



Graphic Analysis: The Method of Successive Elimination

Let us consider the problem of a certain real estate broker who has purchased a tract of land for subdivision into lots. He wished to know how much the area and the view from these lots contributed to their value. He also wanted a method for setting a reasonable price on the lots.

In order to obtain some information, the broker selected 20 nearby lots that had been recently sold. He obtained the sale price for each lot and its size (in thousands of square feet). Since he knew the lots at higher altitudes had more value because of the view, he also estimated the elevation of each lot (in feet above sea level). The data are presented in Table 23-3.

Scatter diagrams showing the relationships between each pair of variables are displayed in Chart 23-2. We see that there is a positive linear correlation between price and area and between price and eleva-

tion, but there is no apparent relationship between elevation and area for the 20 lots selected.

The first step in the graphic approach (called the "method of successive elimination") is to determine the simple regression line between the dependent variable Y (price) and the independent variable that is deemed most important. We shall select the area (X_1). This line can be determined by either graphic or least squares techniques, as described in Chapter 22. The equation is $Y_e = 1.45 + 0.219 X_1$ and is shown in Chart 23-3. The slope of the line indicates that the price of a lot increases \$219, on the average, for every thousand square feet of area. This equation, of course, does not take the elevation of the lot into account.

The next step is to eliminate the effect of area on the price of each lot. This is done by subtracting 0.219 for each thousand square feet from the price of the lot. This adjustment to a "no area" basis may be done graphically by measuring the vertical deviations from the regression line in Chart 23-3, or it may be done arithmetically as shown in Table 23-4.

The new price Y' (where $Y' = Y - 0.219X_1$) represents the price adjusted for differences in the size of the lots. This adjusted price is then plotted against the second independent variable, elevation (X_2), as shown in Chart 23-4.

Note that the adjustment of price for the effect of the size of the lots considerably improved the relationship between price and elevation. (Compare Chart 23-4 with Chart 23-2B.) The regression line between *adjusted* price and elevation is $Y'_e = -4.09 + 0.0317X_2$. This indi-

Chart 23-3

REGRESSION LINE BETWEEN PRICE AND AREA

REGRESSION EQUATION: $Y_e = 1.45 + 0.219X_1$

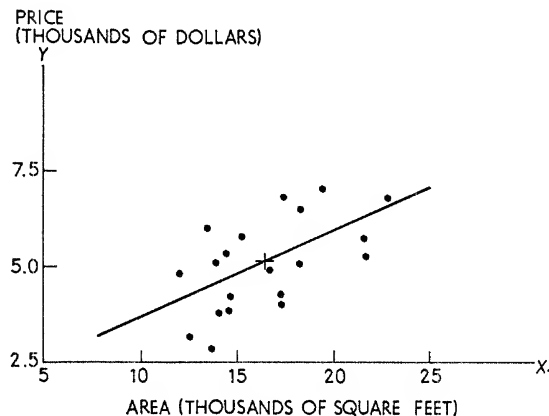


Table 23-4

ADJUSTING PRICE OF LOTS FOR EFFECT OF AREA

Lot No.	X_1 Area, Thousands of Square Feet	Adjustment for Area, $0.219 \times X_1$	Y Price, Thousands of Dollars	$Y' = Y - .219X_1$ Adjusted Price, Thousands of Dollars
1	14.7	3.22	4.1	0.88
2	14.2	3.11	3.9	0.79
3	12.7	2.78	3.2	0.42
4	13.8	3.02	2.9	-0.12
5	14.4	3.15	3.9	0.75
6	17.4	3.81	4.1	0.29
7	21.8	4.77	5.8	1.03
8	14.0	3.07	5.1	2.03
9	17.5	3.83	6.8	2.97
10	23.0	5.04	6.8	1.76
11	18.3	4.01	6.5	2.49
12	19.4	4.25	7.0	2.75
13	15.2	3.33	5.8	2.47
14	18.3	4.01	5.1	1.09
15	21.7	4.75	5.3	0.55
16	16.7	3.66	4.9	1.24
17	13.6	2.98	6.0	3.02
18	14.5	3.18	5.3	2.12
19	12.1	2.65	4.8	2.15
20	17.4	3.81	4.3	0.49
			Total	29.17
			Average	1.4585

cates that the price of a lot increases about \$32 for every foot of elevation—after eliminating the effect of area on price.

We can include the effect of both area and elevation in one equation by taking the term of the first equation that shows the increase in price per unit increase in area and adding it to the second equation, as follows: $Y_c = -4.09 + 0.219X_1 + 0.0317X_2$. This is a first approximation to the multiple regression equation.¹

To refine the estimate, the original price should be adjusted for the effect of elevation (by subtracting 0.0317 for each foot of elevation).

¹ In this case, the first approximation is very close to the least squares equation $Y_c = -3.86 + 0.203X_1 + 0.0319X_2$. This is because X_1 and X_2 are uncorrelated. If X_1 and X_2 were highly correlated, a number of successive approximations would be necessary before the graphic fit converged on the least squares equation. See M. Ezekiel and K. A. Fox, *Methods of Correlation and Regression Analysis*, 3d ed. (New York: John Wiley, 1959), Chap. 10.

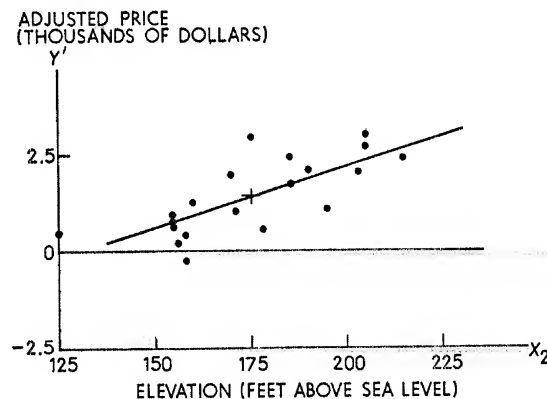
The resulting adjusted price would then be plotted against area (X_1) to obtain a more refined estimate of the net regression coefficient b_1 . After this step, the value of b_2 could be refined, using the improved relationship between Y and X_1 . The process could then be repeated until stable values are obtained for b_1 and b_2 .

Little value can be achieved by following this process further. Our object is merely to describe the graphic method in multiple regression and to clarify the meaning of the net regression coefficient. One can see

Chart 23-4

REGRESSION LINE BETWEEN ADJUSTED PRICE AND ELEVATION

$$\text{REGRESSION EQUATION: } Y'_c = -4.09 + 0.0317 X_2$$



from this analysis how the value of the net regression coefficient depends upon the other variables in the regression equation.

Finding the Regression Equation by Least Squares

Just as in the case of simple regression analysis, the constants of the linear multiple regression equation are determined by the method of least squares by solving a system of simultaneous linear equations, called the *normal equations*, in which the unknowns are the constants of the regression equation. In order to find the constants in the three-variable linear multiple regression

$$Y_c = a + b_1X_1 + b_2X_2$$

the following three normal equations must be solved:

$$\begin{aligned}\Sigma Y &= na + b_1\Sigma X_1 + b_2\Sigma X_2 \\ \Sigma X_1Y &= a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \\ \Sigma X_2Y &= a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2\end{aligned}$$

These equations can be solved directly, but it is usually simpler to measure each variable as a deviation from its mean, as we did in simple regression. That is, we use small x 's and y 's, where $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, and $y = Y - \bar{Y}$. This is done most easily by totaling the squares and products of the original X 's and Y 's, as called for in the above formulas, and then subtracting the *mean times the sum* of the respective variables to get the sums of the small x 's and y 's as follows:

$$\begin{array}{lcl} \frac{\Sigma X_1^2}{-\bar{X}_1 \Sigma X_1} & \frac{\Sigma X_2^2}{-\bar{X}_2 \Sigma X_2} & \frac{\Sigma Y^2}{-\bar{Y} \Sigma Y} \\ = \Sigma x_1^2 & = \Sigma x_2^2 & = \Sigma y^2 \end{array} \quad \begin{array}{lcl} \frac{\Sigma X_1 Y}{-\bar{X}_1 \Sigma Y} & \frac{\Sigma X_2 Y}{-\bar{X}_2 \Sigma Y} & \frac{\Sigma X_1 X_2}{-\bar{X}_1 \Sigma X_2} \\ = \Sigma x_1 y & = \Sigma x_2 y & = \Sigma x_1 x_2 \end{array}$$

The calculation of the adjusted sums of squares and cross products is shown in Table 23-5 for our example of the price of residential lots.

Table 23-5

MULTIPLE REGRESSION BETWEEN AREA (X_1), ELEVATION (X_2), AND PRICE (Y) OF 20 LOTS

CALCULATION OF ADJUSTED SUMS OF SQUARES AND CROSS PRODUCTS

Sum of Variable Mean Adjustment (Mean Times Sum)	Symbols								
	ΣX_1 \bar{X}_1	ΣX_2 \bar{X}_2	ΣY \bar{Y}	ΣX_1^2	ΣX_2^2	ΣY^2	$\Sigma X_1 Y$	$\Sigma X_2 Y$	$\Sigma X_1 X_2$
				$-\bar{X}_1 \Sigma X_1$ Σx_1^2	$-\bar{X}_2 \Sigma X_2$ Σx_2^2	$-\bar{Y} \Sigma Y$ Σy^2	$-\bar{X}_1 \Sigma Y$ $\Sigma x_1 y$	$-\bar{X}_2 \Sigma Y$ $\Sigma x_2 y$	$-\bar{X}_1 \Sigma X_2$ $\Sigma x_1 x_2$
Which Gives									
<i>Residential Lot Example</i>									
Sum	330.7	3501.	101.6	5,657.41	622,729	543.440	1,721.480	18,119.90	57,985.3
Mean	16.535	175.05	5.08						
Adjustment (Mean Times Sum)				-5,468.12	-612,850	-516.128	-1,679.956	-17,785.08	-57,889.0
Adjusted Total				189.29	9,879.	27.312	41.524	334.82	96.3

SOURCE: Table 23-3.

The individual squares and products are not shown because they are usually cumulated in a calculating machine and only the totals need be recorded.²

² Since the normal equations for a three-variable problem involve quite a number of sums of squares and products, it is important to choose a system of internal checks, when using hand calculators. In this connection a sum variable,

$$X_s = X_1 + X_2 + Y$$

is extremely useful. In addition to the comparatively simple check,

$$\Sigma X_s = \Sigma X_1 + \Sigma X_2 + \Sigma Y$$

the sum of squares of X_s provides the check

$$\Sigma X_s^2 = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma Y^2 + 2\Sigma X_1 Y + 2\Sigma X_2 Y + 2\Sigma X_1 X_2$$

When we express the second and third normal equations in small x 's, the terms Σx_1 and Σx_2 equal zero, and the equations become

$$\begin{aligned}\Sigma x_1 y &= b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 \\ \Sigma x_2 y &= b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2\end{aligned}$$

Substituting the numerical values from Table 23-5, we have

$$\begin{aligned}41.524 &= 189.29b_1 + 96.3b_2 \\ 334.82 &= 96.3b_1 + 9,879.b_2\end{aligned}$$

These equations can be solved simultaneously to find b_1 and b_2 as follows: Multiply the first equation by $96.3/189.29$, the ratio of the b_1 coefficients. The result is

$$21.225 = 96.3b_1 + 48.992b_2$$

Subtract this from the second normal equation to eliminate b_1 . Then,

$$313.695 = 9,830.0b_2$$

and

$$b_2 = 0.03191$$

Substitute this value of b_2 in the first normal equation. Solving,

$$b_1 = 0.2031$$

Finally, substitute both values in the second equation as a check on the arithmetic.

The value of the constant a is

$$\begin{aligned}a &= \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \\ &= 5.080 - (0.2031)(16.535) - (0.03191)(175.05) \\ &= -3.864\end{aligned}$$

Now, substitute the three constants in the multiple regression equation

$$\begin{aligned}Y_e &= a + b_1 X_1 + b_2 X_2 \\ &= -3.864 + 0.2031X_1 + 0.03191X_2\end{aligned}$$

Thus, for a lot with 15 thousand square feet ($X_1 = 15.0$) and elevation of 180 feet ($X_2 = 180$), the estimated price would be

$$\begin{aligned}Y_e &= -3.864 + 0.2031(15.0) + 0.03191(180) \\ &= 4.926 \text{ thousands of dollars, or nearly } \$5,000\end{aligned}$$

Standard Error of Estimate

Just as in simple correlation, the standard error of estimate is in effect the standard deviation of the residuals, $Y - Y_e$. It measures the aver-

age scatter of Y values around the regression plane. The standard error of estimate is

$$S_{Y \cdot 12} = \sqrt{\frac{\sum (Y - Y_c)^2}{n - k}}$$

where n is the number of observations and k is the number of constants in the regression equation. Here, $n = 20$ and $k = 3$. The symbol $S_{Y \cdot 12}$ denotes the standard error of estimate of the dependent variable Y regressed against the two independent variables X_1 and X_2 .

It is difficult to calculate $\sum (Y - Y_c)^2$ directly, so we use the following equivalent formula for computation purposes:

$$S_{Y \cdot 12} = \sqrt{\frac{\sum y^2 - b_1 \sum x_1 y - b_2 \sum x_2 y}{n - k}}$$

In our example,

$$\begin{aligned} S_{Y \cdot 12} &= \sqrt{\frac{27.312 - (0.2031)(41.524) - (0.03191)(334.82)}{20 - 3}} \\ &= \sqrt{0.4820} \\ &= 0.694 \text{ or about } \$700 \end{aligned}$$

That is, if prices are normally distributed about the regression plane, about two thirds of the prices should fall within \$700 of the value estimated from the regression equation.

COEFFICIENT OF MULTIPLE DETERMINATION

As in simple correlation, the coefficient of multiple determination is the ratio of explained variance to total variance, or 1 minus the unexplained variance over the total variance. That is,

$$R^2 = 1 - \frac{S_{Y \cdot 12}^2}{s_Y^2}$$

where s_Y^2 is the total variance of the dependent variable Y . In our example, the unexplained variance ($S_{Y \cdot 12}^2$) was found to be 0.4820. The estimated total variance (from Table 23-5) is

$$s_Y^2 = \frac{\sum y^2}{n - 1} = \frac{27.312}{20 - 1} = 1.4375$$

Therefore,

$$R^2 = 1 - \frac{0.4820}{1.4375} = 0.6647$$

About 66 percent of the variance in price, therefore, is explained by the variance in area and elevation of the lots.

The *coefficient of multiple correlation* is the square root of the coefficient of multiple determination. Here,

$$R = \sqrt{0.6647} = 0.815$$

The multiple correlation coefficient is always positive, regardless of the signs of the regression coefficients.

STATISTICAL INFERENCE IN MULTIPLE REGRESSION

When the data used in multiple regression represent a probability sample from some specific population, it is possible to make statistical inferences about the population parameters. In particular, if the population relationship is of the form

$$Y = A + B_1X_1 + B_2X_2 + z$$

where B_1 and B_2 are the "true" net regression coefficients, A is the true intercept, and z is the residual deviation, then the least squares estimates a , b_1 , and b_2 are efficient, linear, unbiased estimates of the corresponding population parameters.

The assumptions underlying this estimation procedure are the same as in simple regression, namely,

1. *Linearity*: For fixed values of X_1 and X_2 , the mean values of Y lie on a rectilinear plane. This implies $E(z) = 0$, where $z = Y - Y_c$.
2. *Independence*: The residuals (z values) are independent of each other.
3. *Uniform Scatter*: The points have a uniform dispersion about the regression plane.
4. *Normality*: The values of z are normally distributed (not necessary for large samples).

Standard Error of the Regression Coefficient

The regression coefficient b_1 is an estimate of the population parameter B_1 . The sampling error associated with this estimate, called the *standard error of the regression coefficient*, for the case of two independent variables (X_1 and X_2) is

$$s_{b_1} = \frac{S_{Y \cdot 12}}{\sqrt{\sum X_1^2 (1 - r_{12}^2)}}$$

where r_{12}^2 is the coefficient of determination between X_1 and X_2 . Similarly,

$$s_{b_2} = \frac{S_{Y \cdot 12}}{\sqrt{\sum x_2^2 (1 - r_{12}^2)}}$$

In our example (ignoring the correction for sample bias),

$$\begin{aligned} r_{12}^2 &= \frac{(\sum x_1 x_2)^2}{(\sum x_1^2)(\sum x_2^2)} = \frac{(96.3)^2}{(189.29)(9,879)} \\ &= .0050 \end{aligned}$$

and the standard errors of the regression coefficients are

$$\begin{aligned} s_{b_1} &= \frac{S_{Y \cdot 12}}{\sqrt{\sum x_1^2 (1 - r_{12}^2)}} = \frac{0.6942}{\sqrt{(189.29)(0.995)}} \\ &= 0.0506 \end{aligned}$$

and

$$\begin{aligned} s_{b_2} &= \frac{S_{Y \cdot 12}}{\sqrt{\sum x_2^2 (1 - r_{12}^2)}} = \frac{0.6942}{\sqrt{(9,879)(0.995)}} \\ &= 0.0070 \end{aligned}$$

We can test the hypothesis that either area or elevation has zero effect (that is, either $B_1 = 0$ or $B_2 = 0$) by computing b_1/s_{b_1} or b_2/s_{b_2} . In the case of B_1 , the sample value of b_1 is $0.2031/0.0506 = 4.01$ standard errors away from zero. And the sample value of b_2 is $0.03191/0.0070 = 4.56$ standard errors from a hypothesized $B_2 = 0$. The t value (Appendix J) with $n - k$ degrees of freedom is used to make this test. Here, $n = 20$ and $k = 3$, the total number of variables, so $n - k = 17$. The two-tailed t value at the 0.01 level of probability is 2.898 for 17 degrees of freedom. Hence, both B_1 and B_2 are significantly different from zero at the 0.01 level.

The *standard error of the regression plane* and the *standard error of forecast* can be calculated for multiple regression just as in simple regression. The reader is referred to Appendix B at the end of this chapter (p. 627) for the calculations.

INTERPRETATION OF MULTIPLE REGRESSION RESULTS

In simple regression, the regression line, the standard error of estimate, and other calculated values were relatively easy to interpret. In multiple regression, the interpretation is more difficult, since we must

sort out the importance of each variable and the interactions between them.

Beta Coefficients

The regression coefficients b_1 , b_2 , etc. measure the net effect of each variable on the dependent variable Y . But since each of the variables X_1 , X_2 , etc. may be in different units (in our example X_1 is in thousands of square feet and X_2 is in feet above sea level), it is difficult to ascertain the relative importance of each X in influencing Y . One means of accomplishing this is by using β (*beta*) coefficients. These are defined to be

$$\begin{aligned}\beta_1 &= b_1 \left(\frac{s_{X_1}}{s_Y} \right) = b_1 \sqrt{\frac{\sum x_1^2}{\sum y^2}} \\ \beta_2 &= b_2 \left(\frac{s_{X_2}}{s_Y} \right) = b_2 \sqrt{\frac{\sum x_2^2}{\sum y^2}} \\ &\text{etc.}\end{aligned}$$

The β coefficients are merely the net regression coefficients adjusted by expressing each variable in units of its own standard deviation. This adjustment eliminates the effects of the different size and type of the variables and puts the regression coefficients on a comparable basis. In our example,

$$\begin{aligned}\beta_1 &= b_1 \sqrt{\frac{\sum x_1^2}{\sum y^2}} = (0.2031) \sqrt{\frac{189.29}{27.312}} \\ &= 0.535\end{aligned}$$

and

$$\begin{aligned}\beta_2 &= b_2 \sqrt{\frac{\sum x_2^2}{\sum y^2}} = (0.03191) \sqrt{\frac{9,879}{27.312}} \\ &= 0.607\end{aligned}$$

That is, for each increase of one standard deviation in X_1 (area), the price increases by 0.535 standard deviations, while for every increase of one standard deviation in X_2 (elevation), the price increases by 0.607 standard deviations. The two betas are pure numbers and are comparable. Therefore, elevation is slightly more important than area in determining the price of a lot.

Use of Computer Programs

In the previous example, the analysis for three variables could be performed easily by hand calculators. With more than three variables,

however, the analysis becomes increasingly complicated, since the number of normal equations to be solved for the linear regression equation increases with the number of independent variables. (We cannot visualize a regression plane, as in Chart 23-1, for more than three dimensions, but we can still consider the regression equation as a hyperplane in any number of dimensions.) One solution is to use matrix methods, as described in Appendixes A and B at the end of this chapter. There are also many multiple regression programs available for electronic computers.

We will here describe a typical computer program—specifically the BMD02R multiple regression program,³ and interpret its printout sheet.

Table 23-6

CHARACTERISTICS AFFECTING THE PRICE OF 20 LOTS

Lot No.	AREA Thousands of Square Feet	ELEVATION Feet Above Sea Level	SLOPE Degrees	VIEW Scale 1 (Poor) to 9 (Excellent)	PRICE Thousands of Dollars
1	14.7	155	1.5	2	4.1
2	14.2	155	1.8	2	3.9
3	12.7	158	2.9	1	3.2
4	13.8	158	1.0	1	2.9
5	14.4	155	0.5	2	3.9
6	17.4	157	1.0	2	4.1
7	21.8	172	5.7	4	5.8
8	14.0	170	5.4	6	5.1
9	17.5	175	17.5	9	6.8
10	23.0	185	14.5	9	6.8
11	18.3	185	14.4	9	6.5
12	19.4	205	12.2	9	7.0
13	15.2	215	5.0	8	5.8
14	18.3	195	13.1	6	5.1
15	21.7	178	15.2	8	5.3
16	16.7	160	10.1	8	4.9
17	13.6	205	7.4	7	6.0
18	14.5	190	5.8	7	5.3
19	12.1	203	5.1	7	4.8
20	17.4	125	17.3	1	4.3
Total	330.7	3501.	157.4	108.	101.6
Mean	16.535	175.05	7.87	5.40	5.08

³ Described in *BMD Biomedical Computer Programs*, Health Services Computing Facility, School of Medicine, University of California, Los Angeles, January 1, 1964, pp. 233-53. The program output is modified to eliminate some detail and certain statistical measures that are not explained in this text.

Table 23-7—Continued

STEP NUMBER 1
VARIABLE ENTERED 5

MULTIPLE R 0.8787
STD. ERROR OF EST. 0.5881

VARIABLES IN EQUATION			VARIABLES NOT IN EQUATION	
VARIABLE	COEFFICIENT	STD. ERROR	VARIABLE	PARTIAL CORR.
(CONSTANT	3.26574)		AREA 2	0.52309
VIEW 5	0.33597	0.04303	ELEVTV 3	-0.04302
			SLOPE 4	0.34439

STEP NUMBER 2
VARIABLE ENTERED 2

MULTIPLE R 0.9135
STD. ERROR OF EST. 0.5158

VARIABLES IN EQUATION			VARIABLES NOT IN EQUATION	
VARIABLE	COEFFICIENT	STD. ERROR	VARIABLE	PARTIAL CORR.
(CONSTANT	1.77976)		ELEVTV 3	0.19185
AREA 2	0.10333	0.04083	SLOPE 4	0.09071
VIEW 5	0.29475	0.04110		

STEP NUMBER 3
VARIABLE ENTERED .3

MULTIPLE R 0.9168
STD. ERROR OF EST. 0.5218

VARIABLES IN EQUATION			VARIABLES NOT IN EQUATION	
VARIABLE	COEFFICIENT	STD. ERROR	VARIABLE	PARTIAL CORR.
(CONSTANT	0.62111)		SLOPE 4	0.21297
AREA 2	0.11629	0.04451		
ELEVTV 3	0.00668	0.00854		
VIEW 5	0.25321	0.06746		

means and standard deviations of each variable. The "covariance matrix" gives the average of the product of each pair of variables, expressed as deviations from their means. Thus, $\Sigma x_1 x_2 / n = 2.185$. The items on the diagonal are variances, for example, $\Sigma x_1^2 / n = 1.437$, the square of the standard deviation of X_1 , which is 1.19895.⁴

The "correlation matrix" shows the coefficient of simple correlation between each pair of variables. Note that all the variables are positively related to the dependent variable—price—with correlation coefficients ranging from 0.578 to 0.879.

⁴ The standard deviations, variances, and correlation coefficients in this program are sample values, not adjusted for degrees of freedom.

Table 23-7—Continued

STEP NUMBER 4
 VARIABLE ENTERED 4
 MULTIPLE R 0.9207
 STD. ERROR OF EST. 0.5265

VARIABLE	VARIABLES IN EQUATION		VARIABLES NOT IN EQUATION	
	COEFFICIENT	STD. ERROR	VARIABLE	PARTIAL CORR.
(CONSTANT	0.24021			
AREA 2	0.09873	0.04950		
ELEVTV 3	0.01068	0.00983		
SLOPE 4	0.02950	0.03494		
VIEW 5	0.20487	0.08896		

SUMMARY TABLE

STEP NUMBER	VARIABLE ENTERED REMOVED		MULTIPLE R	RSQ	INCREASE IN RSQ
1	VIEW	5	0.8787	0.7720	0.7720
2	AREA	2	0.9135	0.8344	0.0624
3	ELEVTV	3	0.9168	0.8405	0.0061
4	SLOPE	4	0.9207	0.8477	0.0072

LIST OF RESIDUALS

CASE	RESIDUAL		
1	0.29968	11	0.20937
2	0.14019	12	0.45214
3	-0.27132	13	-0.02269
4	-0.62388	14	-0.64444
5	0.15879	15	-1.07031
6	0.02650	16	-0.63405
7	0.58357	17	0.57611
8	0.27414	18	-0.00541
9	0.60367	19	-0.38660
10	0.04239	20	0.29218

In the stepwise procedure, the program first calculates the simple regression between price and the independent variable that explains the greatest part of the variation in price (the dependent variable). In this case the variable "view" (number 5) is first included, since $r_{15} = 0.879$ —the highest value in the top row of the correlation matrix. The next lines show this value, the standard error of estimate, the coefficients a and b_5 , and the standard error of the latter.

In the next step, a second independent variable is included in the regression. The factor chosen is the one that makes the greatest additional contribution to explained variance. The right-hand column labeled "Partial Correlation" or *partial correlation coefficient* gives an indication at each stage of the relative importance of each of the variables not yet in the regression equation. The square of the partial correlation coefficient measures the increase in explained variance from the addition of a given variable *relative to* the variance remaining to be

explained before the variable was added. Thus, the partial correlation coefficient indicates which variable would have the greatest effect (in reducing unexplained variance) if added to the regression. In this step, the variable "area" (number 2) is added, increasing the multiple correlation coefficient to 0.9135.

Variables 3 and 4 (elevation and slope) are added in turn but have little effect on the multiple correlation coefficient. At the end of step 4, all variables are included in the regression equation. A summary table is printed showing the cumulative correlation coefficient R , as well as R^2 and the increase in R^2 caused by the introduction of each variable.

The "List of Residuals" gives the variation in price of each lot not explained by the multiple regression equation. As an optional feature, the computer will plot these residual terms against each of the independent variables. Such a plot is shown in Chart 23-5 for variable 2 (area) and is a useful check on the assumptions of linearity and homoscedasticity. The scatter seems approximately uniform over the range of the independent variable, and there is no evidence of curvilinearity. (The same is true of the other three plots, not shown.) Hence, we can conclude that the linear and homoscedasticity assumptions are satisfied (though the sample size of 20 is too small for us to be certain).

Tests of Significance. The inclusion of the standard errors of the net regression coefficients makes it possible to test for their significance. In particular, we can test whether each coefficient is significantly different from zero. The test is performed using the t value (Appendix J) with $(n - k)$ degrees of freedom, where k is the number of variables. For $20 - 5 = 15$ degrees of freedom, the two-tailed t value at the 0.05 level is 2.131. The variable "view" is significant at this level since the regression coefficient is 2.30 standard errors ($0.20487/0.08896 = 2.30$) from zero. And "area" is nearly significant ($0.09873/0.04950 = 1.99$). However, neither "elevation" nor "slope" is close to significance at the 0.05 level (for elevation, $0.01068/0.00983 = 1.10$; for slope, $0.02950/0.03494 = 0.844$). Hence, we might well discard these factors and express price as a function of just area and view (Table 23-7):

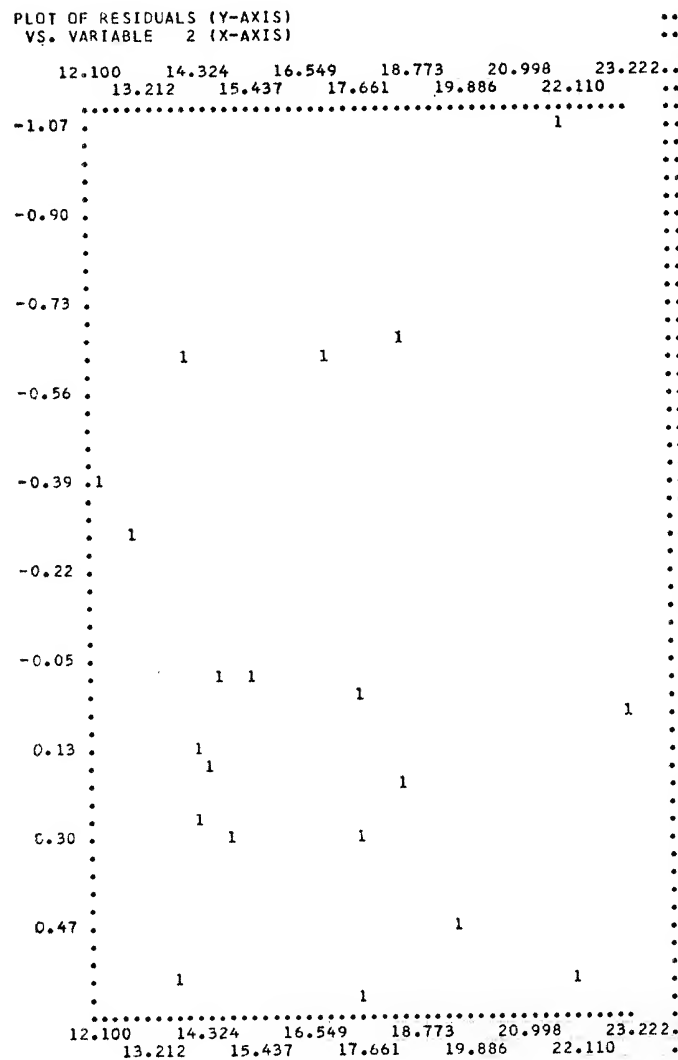
$$\text{Price} = 1.77976 + 0.10333 \times \text{area} + 0.29475 \times \text{view}$$

CAUTIONS IN THE USE OF MULTIPLE REGRESSION

Basic Assumptions

The use of multiple regression formulas in making inferences implies the assumptions that the residuals $z = Y - Y_e$ are (1) clustered around

Chart 23-5



a rectilinear (not curved) plane, (2) independent of each other (3) uniform in their scatter, and, for small samples, (4) normally distributed. If these assumptions are not valid, conclusions from multiple regression analysis may be very misleading. Yet they are often overlooked because of the ease in running a computer program and the difficulty of checking the assumptions mathematically. A simple graphic check is to first plot the original variables against each other, as in Chart 23-2, and then, after running the program, to plot the residuals against

each independent variable, as in Chart 23-5. The residuals can then be checked visually for these conditions.

The same distinction should be made between the regression model and the correlation model as in simple correlation (see Chapter 22).

A second major source of error in using regression analysis is to extrapolate beyond the range of the data upon which the regression equation was estimated. The equation by itself gives no indication of what lies outside the range of its data—the surface may become curvilinear, for example. Nevertheless, it is sometimes necessary to extrapolate, such as when we make economic forecasts, or apply a relationship for one region to another comparable region. For such a projection to be valid, it is essential that the pertinent economic conditions in the extrapolated period or region be essentially similar to those on which the regression analysis was based.

Colinearity

When the independent variables in a multiple regression are highly correlated with each other, the net regression coefficients may be unreliable. This can be seen easily from the formula for the standard error of the regression coefficient in the case of two independent variables:

$$s_{b_1} = \frac{S_{Y \cdot 12}}{\sqrt{\sum X_1^2 (1 - r_{12}^2)}}$$

where r_{12} is the correlation coefficient between the independent variables X_1 and X_2 .

The standard error is smallest when r_{12} is zero, but as r_{12} approaches one (perfect correlation), the denominator of the equation approaches zero, and the standard error becomes very large, so the regression coefficient itself becomes unreliable. Hence, the standard error is sensitive to the *colinearity* or correlation between X_1 and X_2 . This accords with common sense: If X_1 and X_2 move together, it is difficult to distinguish their separate effects on Y . One solution is simply to drop the X that is deemed less important.⁵

While colinearity affects the reliability of *individual* variables in the regression, it may not alter the predictive power of the *total* regression

⁵ The effects of colinearity may be seen in the computer regression example (Table 23-7). The correlation between elevation (X_3) and view (X_5) is 0.749 and between slope (X_4) and view (X_5) is 0.608. Note what happens to the standard error of X_5 as these other two variables are entered in the regression equation. In step 3, the standard error of X_5 increases from 0.041 to 0.067 as X_3 is included, and further increases to 0.089 as X_4 is also included in step 4.

equation. That is, the standard error of estimate may not be increased. The sampling errors of the regression coefficients tend to compensate for each other in the estimate of the dependent variable. Similarly, the sampling error of the multiple correlation coefficient is not sensitive to colinearity among the independent variables.

Colinearity may produce some peculiar results in regression analysis besides its effect upon the sampling error of the net regression coefficients. For example, two variables X_1 and X_2 may be highly positively correlated with Y and with each other. But the net effect of X_2 , taking X_1 into account, may be negative. For example, on a certain railroad, the number of miles traveled by empty cars may be positively correlated with profit. However, empty-car mileage is highly correlated with full-car mileage. So, when the latter variable is included in the regression equation, the net effect of hauling empty cars may be negative.

SUMMARY

Multiple regression measures the simultaneous influence of a number of independent variables upon one dependent variable. A *net regression coefficient* (e.g., b_1) measures the effect upon the dependent variable of a unit increase in an independent variable, holding the other independent variables constant. The regression equation represents a plane in three-dimensional space or a hyperplane in more than three dimensions.

The multiple regression equation can be estimated either *graphically* or by *least squares*. The graphic method allows for the successive elimination of the effects of one variable at a time and the recursive refinement of the estimate of the regression coefficients.

The least squares method can be performed on a hand calculator for three variables, but for more variables it is preferable to use matrix methods (described in the appendixes of the chapter) or an electronic computer program. To calculate the least squares equation, a set of normal equations must be solved. To make this easier, the sums of the squares and cross products of the variables are adjusted by subtracting the mean times the sum of the appropriate variables to reduce them to deviations from their means.

The *standard error of estimate* is essentially the standard deviation of the residuals $z = Y - Y_c$ about the regression plane. And the *coefficient of multiple determination* is the proportion of the variance of the dependent variable explained by the independent variables. Its square root is the *coefficient of multiple correlation*. These concepts are equivalent to those in simple correlation.

When the assumptions of *linearity, uniform scatter, independence,*

and *normality* are satisfied, it is possible to measure the sampling error of the net regression coefficients. These measures can then be used to make statistical inferences about the true regression relationships.

The net regression coefficients can be expressed in common standard-deviation units by multiplying each one by the standard deviation of the appropriate independent variable over the standard deviation of the dependent variable. These β coefficients may be compared for different independent variables, revealing the relative importance of each variable in the regression equation.

Electronic computer programs are widely available for multiple regression analysis; a typical program is described.

Before using multiple regression results, it is important to check the assumptions upon which the analysis is based. Plots of the original variables and the final residuals versus the independent variables provide a graphic check on these assumptions.

Colinearity or correlation between independent variables reduces the reliability of the net regression coefficients, but it may not affect the predictability of the overall regression equation.

PROBLEMS

1. Suppose we have estimated the least squares linear regression of Y on X_1 and X_2 to be $Y_o = a + b_1X_1 + b_2X_2$. For each of the statements below, indicate in a few sentences why you agree or disagree with the statement.
 - a) If b_1 is 18 times as large as b_2 , then we may infer that X_1 is considerably more important than X_2 in accounting for the variation in Y .
 - b) The number b_1 is intended to measure the expected change in Y in response to a unit change in X_1 with X_2 held constant.
For all of the remaining statements, suppose further that R^2 is very high, say $R^2 = 0.98$.
 - c) The numbers a , b_1 , and b_2 are all estimated to be significantly different from zero.
 - d) The estimated relationship is a very close approximation to the true relationship between Y and X_1 , X_2 .
 - e) The observed Y 's do not vary much from the calculated Y 's.
 - f) Variations in X_1 and X_2 account for a very considerable proportion of the observed variations in Y .
 - g) The observed residuals ($z = Y - Y_o$) show no systematic pattern.
 - h) Dropping either X_1 or X_2 and estimating the simple regression of Y on the remaining variable would not reduce R^2 very much.
 2. In a study of the demand for automobiles, the following regression model was used: $Y_o = a + b_1X_1 + b_2X_2 + b_3X_3$, where Y is expenditures (in billions of dollars) on new cars during year t (the period covered was 1948–1961); X_1 is the price index for all cars, new and used, during
-

period t ; X_2 is the estimated value of the total stock of automobiles at the end of year $t - 1$, in billions of dollars; and X_3 is the per capita disposable income during year t (in dollars).

The following results were obtained from the data:

$$Y_c = 0.0779 - 0.0201X_1 - 0.2310X_2 + 0.0117X_3$$

$$\quad \quad \quad [0.0026] \quad [0.0472] \quad [0.0011]$$

$$R^2 = 0.858$$

where the numbers in the brackets are the standard errors of the respective regression coefficients.

For each of the statements below, indicate briefly why you agree or disagree with the statement.

- a) Price has a more important effect on expenditures for new cars than does per capita disposable income.
- b) If price increased one index point in a given year, other things being equal, expenditures for new cars would decline by \$0.0201 billion, on the average.
- c) Price does not have an important influence on expenditure for new cars.
- d) About 14 percent of the variance in expenditures for new cars must be explained by variables other than stock of automobiles, price, and per capita disposable personal income.
- e) The squares of the simple correlation coefficients between Y and the other variables X_1 , X_2 , and, X_3 , respectively, must equal 0.858, that is, $r^2_{Y.1} + r^2_{Y.2} + r^2_{Y.3} = 0.858$.
- f) The fact that the coefficient of X_2 is approximately ten times as large as the coefficient of X_1 means that X_2 explains considerably more of the variability in Y than does X_1 .
- g) The residuals ($z = Y - Y_c$) are necessarily independent of each other.

3. Annual sales of the ABC Company in millions of dollars (Y) correlate with U.S. disposable personal income in billions (X_1) and company advertising expenditures in millions (X_2), as follows, for 1948–1967:

$$Y_c = 210 + 18X_1 \text{ (simple regression)}$$

$$Y_c = 175 + 6X_1 + 11X_2 \text{ (multiple regression)}$$

- a) What factors are likely to have caused the change in the coefficient of disposable income (X_1) from 18 in the first equation to 6 in the second?
 - b) If advertising expenditures were to be the same next year as this year (i.e., X_2 held constant), would you expect sales to increase \$18 or \$6 million in response to a \$1 billion increase in disposable income? Explain.
4. The personnel director of the Acme Insurance Company wishes to determine whether the selling ability of salesmen can be predicted from their education and age. If so, these criteria would provide a valuable aid in selecting the most promising candidates for employment. As a start, ten salesmen are selected at random and are rated by their supervisor as to sales ability, education, and age. The rating on sales ability covers a seven-point scale, from "Poor" (0) to "Excellent" (6). The education scale varies from

"Did not finish high school" (0) to "Has master's degree" (4). The age scale extends from "Age 20-29" (0) to "Age 60-69" (4). The results are shown below.

Salesman	Sales Ability Y	Education X_1	Age X_2
A	1	0	3
B	1	1	4
C	1	0	2
D	2	2	4
E	2	1	3
F	3	3	1
G	4	2	0
H	4	4	2
I	6	3	0
J	6	4	1
Sum	30	20	20

- a) Compute the multiple linear regression equation by the method of least squares to estimate sales ability from education and age. Show all computations.
 - b) What is the meaning of the net regression coefficient b_1 in this particular case? How would this value differ in meaning from the regression coefficient in simple correlation between sales ability and education alone?
 - c) How would the reliability of b_1 be affected if the younger men generally had more education than the older men?
5. a) Compute the standard error of estimate in Problem 4, and interpret its meaning as applied to predicting the sales ability of future salesmen.
 - b) Compute the coefficient of multiple determination and interpret its meaning in describing the relationship between sales ability, education, and age for salesmen of this type.
6. The supervisor at Acme Insurance Company (Problems 4 and 5) has been seen dating employee K , an attractive brunette. Is his high rating (6.5) of her apparently attributable to favoritism, or can it be reasonably explained by her education ($X_1 = 4$) and her youth ($X_2 = 1$)? Explain your answer.
 7. Hony Pharmacy operates a chain of retail drug stores. As a means of measuring the efficiency of various stores, the management is studying the relationship between the number of employees, the size of the store, and the average daily sales volume for last year. The data can be summarized as follows:

Y = average daily sales for each store in hundreds of dollars
 X_1 = number of employees for each store
 X_2 = size of each store in hundreds of square feet
 $n = 103$ = number of Hony stores

The raw data and the necessary adjustments are summarized in the table.

	Y	X ₁	X ₂	Y ²	X ₁ ²	X ₂ ²	YX ₁	YX ₂	X ₁ X ₂
Total	515	168	824	3,975	5,708	9,092	4,090	5,620	5,944
Mean	5.0	6.0	8.0						
Less adjustment				<u>2,575</u>	<u>3,708</u>	<u>6,592</u>	<u>3,090</u>	<u>4,120</u>	<u>4,944</u>
Adjusted total				1,400	2,000	2,500	1,000	1,500	1,000
Which is				Σy^2	Σx_1^2	Σx_2^2	Σyx_1	Σyx_2	Σx_1x_2

- Estimate the linear regression equation $Y_c = a + b_1X_1 + b_2X_2$, which predicts monthly sales as a function of the number of employees and the size of the store.
 - Are you sure that the values obtained for b_1 and b_2 in the above equation are statistically different from zero?
 - Is the regression equation of much use in predicting sales? (Explain your answer.)
 - One of Hony's newer and larger stores occupies 1,600 square feet and employs 10 people. Average daily sales have been \$1,500. Is this "out of line" with the experience of other Hony stores?
8. A manual dexterity test (X_1) and a finger-dexterity test (X_2) were administered to 25 applicants for jobs as aircraft riveters. After these 25 applicants were hired and trained, their performance was measured by the number of rivets set correctly per minute (Y). A multiple regression analysis is to be performed to evaluate the worth of each test in predicting performance of riveters. We have the following:

	Y	X ₁	X ₂	Y ²	X ₁ ²	X ₂ ²	YX ₁	YX ₂	X ₁ X ₂
Total	200	150	125	2,213	1,000	775	1,400	1,225	800
Mean	8	6	5						

- Estimate the linear regression equation, which predicts performance as a function of the two tests.
 - Test the hypothesis that neither test has any predictive value for performance of riveters.
 - Which test do you consider more important in predicting riveting performance?
 - Calculate the multiple correlation coefficient.
 - A new employee scores 9 on the manual dexterity test and 8 on the finger dexterity test. Predict his riveting performance.
9. A study was undertaken at a John Deere farm machinery plant to determine what variables influenced the time taken to handle a piece of flat metal stock to the bump gauge of a punch press. The length and weight of the metal piece were thought to be significant factors. Accordingly, the handling time,

weight, and length of a sample of 25 pieces of metal were recorded and are presented in the table.

HANDLING TIME, WEIGHT, AND LENGTH
OF 25 PIECES OF METAL

Item	Time (0.001 Min.)	Weight (0.1 Lb.)	Length (0.1 In.)
1	30	5	35
2	32	12	46
3	15	15	63
4	30	31	67
5	25	6	70
6	25	8	83
7	42	37	88
8	35	23	104
9	42	30	134
10	30	34	151
11	52	17	153
12	50	53	164
13	45	56	173
14	50	41	191
15	70	84	196
16	64	62	198
17	64	66	204
18	70	66	208
19	80	63	238
20	88	80	295
21	105	154	308
22	85	50	310
23	85	184	319
24	105	186	324
25	84	122	394
Total	1,403	1,485	4,516
Mean	56.12	59.40	180.64

- a) Estimate the linear regression between the handling time and the length and weight of the pieces of metal.
 - b) Are the effects of the length and weight statistically significant?
 - c) Which factor is more important in determining the handling time?
 - d) Calculate the standard error of estimate and the coefficient of multiple determination.
 - e) Plot the residuals to check the assumptions of linearity and homoscedasticity.
10. A small company, whose main product is pajamas, has a work force composed of women who work over sewing machines. The president of the

company is concerned about the high rate of absenteeism and "sickness" among the women workers, especially since it has tended to fluctuate widely from week to week.

The plant manager claims that it is due to excessive overtime, a result of the president's policy of maintaining a nearly constant workforce, and meeting any increase in demand through overtime. His comment is, "You just can't make a woman work more than 45 hours a week."

The president, however, believes that it has been primarily due to the fluctuating attempts to the part of the Garment Workers Union to unionize the women. Absenteeism has been encouraged by the organizers to apply pressure to management. A measure of the activity of the organizers is provided by the number of union-supported complaints appearing in the suggestion box each week.

You decide to try to find out which factor is the more influential in causing the fluctuations in absenteeism. Accordingly, you compile the appropriate data over the past 26 weeks (all figures are in hundreds). Let Y be the number of girl-hours absent in a given week; X_1 be the total number of overtime hours required in that week; and X_2 be the number of union-supported complaints in the suggestion box that week.

The data are summarized in the following *adjusted* sums of squares and cross products (the variables being expressed as deviations from their means):

$$\begin{array}{ll} \Sigma y^2 = 31.0 & \Sigma yx_1 = 6.80 \\ \Sigma x_1^2 = 8.0 & \Sigma yx_2 = 2.86 \\ \Sigma x_2^2 = 2.32 & \Sigma x_1x_2 = 1.60 \end{array}$$

- a) Calculate the net regression coefficients b_1 and b_2 .
- b) Do either of the factors (overtime or union activity) explain the fluctuations in absenteeism? Which factor appears to be the more significant statistically? Explain.

11-14. The data shown in the table below were collected by Peck and Scherer in their study of large-scale research and development projects. The projects represent weapon system developments undertaken primarily for the Department of Defense. The *development cost factor* is the ratio of the actual cost of the development to the original estimate. Thus, project F cost seven times as much as originally estimated. Similarly, the *development time factor* is the ratio of the actual time taken to complete the project to the original time estimate. *State of the art advance* is an index, designed by Peck and Scherer, to measure the degree to which the development advanced the frontiers of knowledge. A project was given a rating close to 100 if it involved substantial innovations in factors such as materials, aerodynamics, and fuels. The factors *importance of time* and *importance of cost* represented an estimate of the relative importance of speed and cost to the management of the development. For example, project E has a value of 100 for the importance of time and zero for cost, indicating an urgent crash program with virtually no constraints on cost.

TIME-COST AND TECHNICAL PERFORMANCE FACTORS FOR THIRTEEN DEVELOPMENT PROJECTS

Project Code	Development Cost Factor	Development Time Factor	State of the Art Advance	Importance of Time	Importance of Cost
A	4.0	1.0	95	70	25
B	3.5	2.3	65	40	40
C	5.0	1.9	92.5	25	30
D	2.0	n.a.*	55	80	20
E	n.a.*	0.7	95	100	0
F	7.0	1.8	90	50	40
G	3.0	1.3	80	90	10
H	2.0	1.0	50	90	40
I	2.4	1.3	85	60	40
J	2.5	1.3	60	75	50
K	0.7	1.0	80	95	10
L	3.0	1.4	60	50	50
M	—	—	95	95	15
Average	3.2	1.36	77	71	28

* Not available.

SOURCE: Merton J. Peck and Frederick M. Scherer, *The Weapons Acquisition Process: An Economic Analysis* (Boston: Division of Research, Graduate School of Business, Harvard University, 1962), Tables 10.1 and 16.1.

11. *a)* Is the development cost factor (dependent variable) related to the state of the art advance (considering these two factors only)? Is this relationship significant?
 - b)* Now include also the importance of time as another independent variable in the above relationship. Does this considerably improve the relationship? Does each independent variable have a statistically significant effect upon the development cost factor?
 - c)* Compare the regression coefficients of the variable state of the art advance in parts *a* and *b*. How do you explain the difference?
12. *Note:* This problem requires the use of the matrix solution explained in Appendix B at the end of this chapter. Alternatively, the student may use a computer program, if available. Using the data above:
 - a)* Estimate the multiple regression equation relating the development cost factor (dependent variable) to the three independent variables, the state of the art advance, the importance of time, and the importance of cost.
 - b)* Which net regression coefficients are statistically significant?
 - c)* Does the addition of the third variable—importance of cost—improve the explanation of variations in cost overruns?
 - d)* Compare the net regression coefficients obtained in Problem 12(*a*) with those obtained in Problem 11(*b*).
 13. *a)* How much of the variance in the development *time* factor is explained by the variable importance of time (considering these two factors only)?
 - b)* How much of the variance in development time factor is explained by both importance of time and the state of the art advance?
 - c)* Which of the net regression coefficients in part *b* are statistically significant?

14. *Note:* This problem requires the use of the matrix solution explained in Appendix B at the end of this chapter. Alternatively, the student may use a computer program, if available. For the data above:
- Estimate the multiple regression equation between the development time factor (dependent variable) and the independent variables of importance of cost, the state of the art advance, and the importance of time.
 - Which of the net regression coefficients are statistically significant?
15. An analyst for a manufacturing firm wished to explain the variations that had occurred from period to period in the manufacturing cost per unit of the firm's product. Accordingly, he collected the data over the last 20 quarters. He knew that raw material prices and labor costs had varied considerably over the period, and he estimated an index of these costs. Also, the production rate had fluctuated widely in response to customer demand and inventories. The production level for each period was measured as a percent of rated capacity. The data are shown in the table.

Period	Average Manufacturing Cost per Unit	Production Level as a Percent of Rated Capacity	Index of Raw Material and Labor Costs
1	\$3.65	85	80
2	4.22	78	93
3	4.29	82	107
4	5.43	64	115
5	6.62	50	130
6	5.71	62	128
7	5.09	70	116
8	3.99	90	92
9	4.08	94	94
10	4.38	100	110
11	4.28	104	115
12	4.42	82	117
13	5.11	75	128
14	4.88	84	134
15	4.99	86	135
16	4.57	90	135
17	4.84	94	139
18	5.16	80	142
19	5.67	72	147
20	6.26	60	150
Mean	\$4.882	80.10	120.35

- Determine the multiple regression equation relating cost per unit to production level and raw material cost.
- Explain the meaning of the coefficients in the regression equation.
- How well do these factors explain or predict cost per unit?
- Plot the residuals ($Y - Y_c$) against the independent variables. Is there any evidence of curvilinearity from these plots?

- e) For next quarter, the raw materials and labor cost index is expected to drop to 145, and the production level is expected to rise to 80 percent of capacity. What average manufacturing cost per unit would you expect? Should you qualify your estimate as a result of your answer to part *d* above?

SELECTED READINGS

Selected readings for this chapter are included in the list which appears on page 657.

APPENDIX A: INTRODUCTION TO MATRIX OPERATIONS

Definition of a Matrix

A *matrix* is a rectangular array of elements (numbers or symbols). An example of a matrix, denoted by the symbol A , is shown below:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$

This matrix is the array of the symbols a_{11} through a_{34} . It has three rows and four columns. Each symbol a_{ij} refers to the element in the i th row and the j th column. A matrix is *rectangular*, indicating that it has the same number of elements in each row and in each column (although the number of rows may not equal the number of columns).

A matrix with only one row or column is usually called a *vector*. The vector $[a_1, a_2, a_3, \dots, a_n]$ is an example of a *row* vector (one row), and

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix}$$

is an example of a *column* vector.

The number of rows and columns define the *dimensions* of a matrix. A matrix with 3 rows and 4 columns is said to have dimension 3×4 or, more simply, is a 3×4 matrix. A matrix with the same number of rows and columns is a *square* matrix.

Addition and Subtraction of Matrices

Two matrices may be added (or subtracted) simply by adding (or subtracting) the corresponding elements on an element-by-element basis. However, in order to add (or subtract) the matrices, they must be of the same dimension.

As an example, consider the matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

The sum $A + B$ is defined to be

$$\begin{aligned} A + B &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \end{bmatrix} \end{aligned}$$

That is, the element in the first row and column of A is added to the element in the first row and column of B and so on.

Using an example with numbers, if

$$C = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}$$

then

$$C - D = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

The Transpose of a Matrix

The *transpose* of a matrix A (the transpose is designated A') is obtained by interchanging the rows and columns. Thus, for

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \quad (3 \times 2 \text{ matrix})$$

the transpose

$$A' = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix} \quad (2 \times 3 \text{ matrix})$$

Using a numerical example, if

$$B = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \quad \text{then} \quad B' = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$$

The use of the transpose operation converts a row vector into a column vector and vice versa.

Matrix Multiplication

Matrices may also be multiplied. The rules for matrix multiplication, however, are more complicated than matrix addition. Consider the matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

The product $A \times B$ is

$$\begin{aligned} A \times B &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \\ &= \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}) & (a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32}) \\ (a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31}) & (a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32}) \end{bmatrix} \end{aligned}$$

That is, the element in the *first* row, *first* column, of the product matrix ($A \times B$) is obtained by multiplying and then summing the elements of the *first* row in A and the *first* column in B ; the element in the *first* row, *second* column, of the product matrix ($A \times B$) is obtained by multiplying and then summing the elements of the *first* row of A and the *second* column of B ; the element in the *second* row, *first* column, of ($A \times B$) is obtained by multiplying and then summing the elements of the *second* row of A and the *first* column of B ; and so on.

A numerical example will help to illustrate matrix multiplication:

$$\begin{aligned} A &= \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} \quad B = \begin{bmatrix} -1 & 1 \\ 3 & -3 \end{bmatrix} \\ (A \times B) &= \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} \times \begin{bmatrix} -1 & 1 \\ 3 & -3 \end{bmatrix} \\ &= \begin{bmatrix} (2 \cdot (-1) + 4 \cdot 3 = 10) & (2 \cdot 1 + 4 \cdot (-3) = -10) \\ (6 \cdot (-1) + 8 \cdot 3 = 18) & (6 \cdot 1 + 8 \cdot (-3) = -18) \end{bmatrix} \end{aligned}$$

or

$$(A \times B) = \begin{bmatrix} 10 & -10 \\ 18 & -18 \end{bmatrix}$$

As another example,

$$C = \begin{bmatrix} 5 & 3 \\ 2 & -1 \\ 1 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 2 & 1 & 2 \\ 5 & 4 & 6 \end{bmatrix}$$

$$\begin{aligned}
 C \times D &= \begin{bmatrix} (5 \cdot 2 + 3 \cdot 5 = 25) & (5 \cdot 1 + 3 \cdot 4 = 17) \\ (2 \cdot 2 + (-1) \cdot 5 = -1) & (2 \cdot 1 + (-1) \cdot 4 = -2) \\ (1 \cdot 2 + 0 \cdot 5 = 2) & (1 \cdot 1 + 0 \cdot 4 = 1) \end{bmatrix} \\
 &\quad \begin{bmatrix} (5 \cdot 2 + 3 \cdot 6 = 28) \\ (2 \cdot 2 + (-1) \cdot 6 = -2) \\ (1 \cdot 2 + 0 \cdot 6 = 2) \end{bmatrix} \\
 &= \begin{bmatrix} 25 & 17 & 28 \\ -1 & -2 & -2 \\ 2 & 1 & 2 \end{bmatrix}
 \end{aligned}$$

Dimensions. In order to multiply two matrices, the number of *columns* in the *first* matrix must equal the number of *rows* in the *second*. Otherwise, multiplication is not defined. The product matrix has the same number of rows as the first matrix and the same number of columns as the second matrix.

For example, a (2×4) matrix (2 rows, 4 columns) can be multiplied by a (4×3) matrix resulting in a (2×3) matrix

$$\begin{array}{c}
 \text{same} \swarrow \\
 \text{[that is, } (2 \times 4) \times (4 \times 3) \rightarrow (2 \times 3)\text{]}. \\
 \nwarrow \text{define} \quad \uparrow
 \end{array}$$

Note that a (2×4) matrix cannot be multiplied by another (2×4) matrix.

Order of Multiplication. In ordinary multiplication, the order is not important. That is, 5 times 2 gives the same result as 2 times 5. In matrix multiplication, however, the order in which the matrices are multiplied makes a difference. The matrix multiplication $A \times B$ generally does not give the same result as $B \times A$. For example, if

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 3 & 0 \\ 1 & 4 \end{bmatrix}$$

then

$$(A \times B) = \begin{bmatrix} 7 & 4 \\ 1 & 4 \end{bmatrix} \quad \text{but} \quad (B \times A) = \begin{bmatrix} 6 & 3 \\ 2 & 5 \end{bmatrix}$$

Hence, when two matrices are to be multiplied it is important to indicate which matrix is on the left (or is first) and which is on the right (or is second).

The Identity Matrix. The identity matrix is a square matrix containing ones along the diagonal and zeros elsewhere. It is usually designated by the symbol I . When the identity matrix is multiplied (either

from the left or right) times another matrix of the same dimensions the result is the original matrix.

For example,

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$(A \times I) = (I \times A) = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = A$$

Matrix Inversion

The inverse of a square matrix A is defined to be a matrix A^{-1} such that

$$A \times A^{-1} = I$$

That is, the product of a matrix times its inverse is the identity matrix I . The inverse for a given matrix may not always exist.⁶ But if it does the inverse A^{-1} may be multiplied by A from either the left or right and will produce the identity matrix. That is,

$$A \times A^{-1} = A^{-1} \times A = I$$

There are several ways to calculate the inverse of a given matrix. We shall present a simple method here without explaining the rationale. The reader is referred to advanced texts for more detail. In general, the calculation of inverses of large matrices (larger than 3×3) is tedious work and should be left to electronic computers.

We start the calculation of the inverse by setting up the matrix to be inverted side by side with the identity matrix. Suppose we wish to invert

$$A = \begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix} \quad \text{We set up} \quad \begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

We can then perform any of the following operations on this set of matrices:

1. Multiply any row by a constant.
2. Add (or subtract) any row from another.
3. Multiply a row by a constant and simultaneously add (or subtract) it from another row (a combination of a and b).

Using the operations 1, 2, and 3, the object is to reduce the set of matrices so that the first is in the form of the identity matrix. The

⁶ A matrix will not have a unique inverse if, for example, two rows are the same. See D. Teichroew, *Introduction to Science in Management* (New York: John Wiley, 1964), chap. 13.

second will then be the desired matrix inverse. That is, we wish to arrive at

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

where $\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$ is the inverse matrix of $\begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix}$, our original matrix.

To accomplish this, we proceed as follows: The original matrices are

$$\begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Step 1: Multiply the first row by $\frac{1}{5}$ (using rule 1). This gives

$$\begin{bmatrix} 1 & \frac{2}{5} \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & 1 \end{bmatrix}$$

Step 2: Subtract row 1 from row 2 (using rule 2). This gives

$$\begin{bmatrix} 1 & \frac{2}{5} \\ 0 & 2\frac{3}{5} \end{bmatrix} \begin{bmatrix} \frac{1}{5} & 0 \\ -\frac{1}{5} & 1 \end{bmatrix}$$

Step 3: Multiply the second row by $1/(2\frac{3}{5})$ or $\frac{5}{13}$ (rule 1). This gives

$$\begin{bmatrix} 1 & \frac{2}{5} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{5} & 0 \\ -\frac{1}{13} & \frac{5}{13} \end{bmatrix}$$

Step 4: Simultaneously multiply row 2 by $\frac{2}{5}$ and subtract it from row 1 (rule 3). This gives

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} (\frac{1}{5} - (-\frac{1}{13})(\frac{2}{5})) & 0 - (\frac{5}{13})(\frac{2}{5}) \\ -\frac{1}{13} & \frac{5}{13} \end{bmatrix}$$

or

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{3}{13} & -\frac{2}{13} \\ -\frac{1}{13} & \frac{5}{13} \end{bmatrix}$$

Hence,

$$\begin{bmatrix} \frac{3}{13} & -\frac{2}{13} \\ -\frac{1}{13} & \frac{5}{13} \end{bmatrix} \text{ is the desired inverse of } \begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix}$$

To verify this result we multiply

$$\begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix} \times \begin{bmatrix} \frac{3}{13} & -\frac{2}{13} \\ -\frac{1}{13} & \frac{5}{13} \end{bmatrix}$$

which gives $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and is a check on our calculations

Solution of Simultaneous Equations Using Matrices

Simultaneous equations may be solved by the use of matrices. For example, suppose we had the following three equations with three unknowns:

$$\begin{aligned} 5x_1 + 2x_2 + x_3 &= 10 \\ 3x_2 + 2x_3 &= 8 \\ 4x_1 + x_3 &= 5 \end{aligned}$$

This set of equations may be expressed in matrix notation as

$$\begin{bmatrix} 5 & 2 & 1 \\ 0 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 8 \\ 5 \end{bmatrix}$$

or letting

$$A = \begin{bmatrix} 5 & 2 & 1 \\ 0 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix}, X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ and } B = \begin{bmatrix} 10 \\ 8 \\ 5 \end{bmatrix}$$

we can write

$$A \times X = B.$$

Multiplying both sides of this equation by A^{-1} (A inverse) we have⁷

$$A^{-1} \times A \times X = A^{-1} \times B$$

But since $A^{-1} \times A = I$, and $I \times X = X$, we have $X = A^{-1} \times B$.

This, in matrix form, is the solution of our equation. All that is needed is A^{-1} , the matrix inverse.

Here the inverse of

$$A = \begin{bmatrix} 5 & 2 & 1 \\ 0 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix}$$

is

$$A^{-1} = \begin{bmatrix} \frac{3}{19} & -\frac{2}{19} & \frac{1}{19} \\ \frac{8}{19} & \frac{1}{19} & -\frac{10}{19} \\ -\frac{12}{19} & \frac{8}{19} & \frac{15}{19} \end{bmatrix}$$

⁷ Care must be taken to multiply from the same side in both cases.

and the product

$$A^{-1} \times B \text{ is } \begin{bmatrix} \frac{3}{19} & -\frac{2}{19} & \frac{1}{19} \\ \frac{8}{19} & \frac{1}{19} & -\frac{10}{19} \\ -\frac{12}{19} & \frac{8}{19} & \frac{15}{19} \end{bmatrix} \times \begin{bmatrix} 10 \\ 8 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Since

$$X = A^{-1} \times B, \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \text{ or } x_1 = 1; x_2 = 2; x_3 = 3$$

This procedure will be applied to regression analysis in Appendix B.

APPENDIX B: MATRIX SOLUTION TO MULTIPLE REGRESSION ANALYSIS

In multiple regression analysis, we must solve the set of normal equations for the values of net regression coefficients. For the case of two independent variables expressed as deviations from their means, the normal equations are

$$\begin{aligned} \Sigma yx_1 &= b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 \\ \Sigma yx_2 &= b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 \end{aligned}$$

This can be written in matrix notation as

$$Y = X \times B \quad \text{where}$$

Y is the vector $\begin{bmatrix} \Sigma yx_1 \\ \Sigma yx_2 \end{bmatrix}$

B is the vector of unknown coefficients $B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

X is the matrix of sums of squares and cross products $\begin{bmatrix} \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_1 x_2 & \Sigma x_2^2 \end{bmatrix}$

In the general case of m independent variables, the normal equations are

$$\begin{aligned} \Sigma yx_1 &= b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 + b_3 \Sigma x_1 x_3 + \cdots + b_m \Sigma x_1 x_m \\ \Sigma yx_2 &= b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 + b_3 \Sigma x_2 x_3 + \cdots + b_m \Sigma x_2 x_m \\ \Sigma yx_3 &= b_1 \Sigma x_1 x_3 + b_2 \Sigma x_2 x_3 + b_3 \Sigma x_3^2 + \cdots + b_m \Sigma x_3 x_m \\ &\vdots \\ \Sigma yx_m &= b_1 \Sigma x_1 x_m + b_2 \Sigma x_2 x_m + b_3 \Sigma x_3 x_m + \cdots + b_m \Sigma x_m^2 \end{aligned}$$

$$\text{Letting } Y = \begin{bmatrix} \Sigma yx_1 \\ \Sigma yx_2 \\ \Sigma yx_3 \\ \vdots \\ \Sigma yx_m \end{bmatrix}, B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix}$$

and

$$X = \begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 & \cdots & \sum x_1 x_m \\ \sum x_1 x_2 & \sum x_2^2 & \sum x_2 x_3 & \cdots & \sum x_2 x_m \\ \sum x_1 x_3 & \sum x_2 x_3 & \sum x_3^2 & \cdots & \sum x_3 x_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_1 x_m & \sum x_2 x_m & \sum x_3 x_m & \cdots & \sum x_m^2 \end{bmatrix}$$

The normal equations are expressed in matrix form, as before, $Y = X \times B$.

To solve this set of equations we need the inverse of the sums of squares and cross products matrix X . And the solution is

$$B = X^{-1} \times Y$$

where X^{-1} is the required inverse.

Example

Using the illustration from page 598 of Chapter 23, the matrix of sums of squares and cross products is

$$X = \begin{bmatrix} 189.29 & 96.3 \\ 96.3 & 9879.0 \end{bmatrix}$$

Using the procedures described in Appendix A, we find the inverse matrix to be

$$X^{-1} = \begin{bmatrix} 0.0053092 & -0.000051754 \\ -0.000051754 & 0.00010173 \end{bmatrix}$$

Multiplying this by the Y vector we have

$$B = X^{-1} \times Y = \begin{bmatrix} 0.0053092 & -0.000051754 \\ -0.000051754 & 0.00010173 \end{bmatrix} \times \begin{bmatrix} 41.524 \\ 334.82 \end{bmatrix}$$

$$\text{or } B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0.2031 \\ 0.03191 \end{bmatrix}$$

or $b_1 = 0.2031$ and $b_2 = 0.03191$ as in the chapter.

Standard Error of Regression Coefficients

We shall first designate the individual elements of the inverse matrix X^{-1} by the symbols c_{ij} . Thus,

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \text{ is the representation of the inverse above}$$

where $c_{11} = 0.0053092$; $c_{12} = c_{21} = -0.000051754$; and $c_{22} = 0.00010173$.

Note that $c_{ij} = c_{ji}$ (here, $c_{12} = c_{21}$). A matrix with this property is called *symmetrical*. Note that both X and X^{-1} are symmetrical.

The standard errors of the net regression coefficients can be estimated as functions of the diagonal elements of the inverse matrix.

In the general case,

$$s_{b_j} = S_{Y \cdot 123 \dots m} \sqrt{c_{jj}}$$

In our example,

$$\begin{aligned} s_{b_1} &= S_{Y \cdot 12} \sqrt{c_{11}} \\ s_{b_2} &= S_{Y \cdot 12} \sqrt{c_{22}} \end{aligned}$$

or

$$s_{b_1} = 0.6942 \sqrt{0.0053092} = 0.0506$$

and

$$s_{b_2} = 0.6942 \sqrt{0.00010173} = 0.0070$$

as in the chapter.

Standard Error of the Regression Plane

The sampling error associated with any point on the regression plane can also be measured. Suppose we are interested in measuring the error of the plane at the point $(X_1, X_2, X_3, \dots, X_m)$. We first measure the distance of this point from the mean of each variable, $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$, etc. The standard error of the regression plane can then be expressed as⁸

$$s_{Y_c} = S_{Y \cdot 123 \dots m} \sqrt{\frac{1}{n} + \sum_{i=1}^m \sum_{j=1}^m c_{ij} x_i x_j}$$

where

$$\begin{aligned} \sum_{i=1}^m \sum_{j=1}^m c_{ij} x_i x_j &= c_{11} x_1^2 + c_{22} x_2^2 + \dots + c_{mm} x_m^2 + 2c_{12} x_1 x_2 + 2c_{13} x_1 x_3 \\ &+ \dots + 2c_{1m} x_1 x_m + 2c_{23} x_2 x_3 + 2c_{24} x_2 x_4 + \dots + 2c_{2m} x_2 x_m + \dots \\ &+ 2c_{(m-1)m} x_{(m-1)} x_m \end{aligned}$$

⁸ This can be expressed simply in matrix notation as

$$s_{Y_c} = S_{Y \cdot 123 \dots m} \sqrt{\frac{1}{n} + z' \times X^{-1} \times z}$$

where $z = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$

and z' is the transpose of z . Note also that $c_{ij} = c_{ji}$ because of the symmetry of both X and X^{-1} .

For our example, let us compute the sampling error of the plane for a point $X_1 = 15.5$ and $X_2 = 165.0$. Since $\bar{X}_1 = 16.535$ and $\bar{X}_2 = 175.05$, $x_1 = 1.035$ and $x_2 = 10.05$.

$$\begin{aligned}
 s_{Y_e} &= S_{Y \cdot 12} \sqrt{\frac{1}{n} + c_{11}x_1^2 + c_{22}x_2^2 + 2c_{12}x_1x_2} \\
 &= 0.6942 \sqrt{\frac{1}{20} + (0.0053092)(1.035)^2 + (0.00010173)(10.05)^2} \\
 &\quad + 2(-0.000051754)(1.035)(10.05) \\
 &= 0.6942 \sqrt{0.0658} = 0.1781
 \end{aligned}$$

Standard Error of Forecast

The *standard error of forecast* is the amount of error associated with making a forecast of a new observation. It includes the standard error of the regression plane plus the scatter about the plane ($S_{Y \cdot 123 \dots m}$). It is estimated for specific values of the independent variables X_1, X_2, \dots, X_m .

The standard error of forecast is

$$s_{Y-Y_e} = \sqrt{S_{Y \cdot 12 \dots m}^2 + s_{Y_e}^2}$$

where s_{Y_e} is the standard error of the regression plane as above.

In our example,

$$\begin{aligned}
 s_{Y-Y_e} &= \sqrt{(0.6942)^2 + (0.1781)^2} \\
 &= \sqrt{0.5137} \\
 &= 0.716 \text{ or } \$716
 \end{aligned}$$

24. CURVILINEAR AND TIME SERIES REGRESSION

THIS CHAPTER treats two topics in regression that are of vital concern to the business or economic analyst. First, many relationships are intrinsically curvilinear; to compute linear regressions is to distort the results. Therefore, we present several simple devices for handling curvilinear regressions. Second, the business economist is often called upon to correlate and forecast time series, such as a company's sales. Time series are not randomly distributed about their regression lines, so that special procedures will be described for their treatment. This is one of the most widespread and controversial applications of regression analysis.

CURVILINEAR REGRESSION

There are frequent situations in which the straight line or rectilinear regression plane would be a very poor fit to the data under analysis. We will suggest three methods of fitting regression curves in such cases: (1) drawing "freehand" curves, together with the method of successive elimination in multiple regression; (2) fitting parabolas or other polynomials by least squares; and (3) transforming the data into logarithms, reciprocals, or other functions so that linear equations can be appropriately applied to these functions.

Graphic Analysis

Simple Regression. Suppose that a fertilizer manufacturer is conducting an experiment to determine the effects of nitrogen fertilizer upon corn yields. He selects 16 fields and has each planted to corn. Four fields receive no nitrogen, four fields receive 40 pounds each, four fields 80 pounds, and four fields 120 pounds. The results of this experiment

are shown in Table 24-1 and Chart 24-1. The average yields for the four groups of fields are listed at the bottom of the table and plotted as circles on the chart. It appears that the four group averages follow a *curved* line, concave downward. This is logical, since increasing amounts of fertilizer may well have successively smaller effects upon corn yield, until some level is reached at which corn yields stabilize or even decline.

A "freehand" regression curve has been drawn through the four group averages in Chart 24-1 with the aid of a French curve, by the method described in Chapter 22. If there were more points scattered

Table 24-1
NITROGEN FERTILIZER AND CORN YIELD
SIXTEEN FIELDS

	Amount of Nitrogen (Pounds)			
	0	40	80	120
Corn Yield (Bushels per Acre)	6	40	72	110
	12	80	112	122
	18	80	112	130
	36	96	128	142
Total Yield	72	296	424	504
Average Yield	18	74	106	126

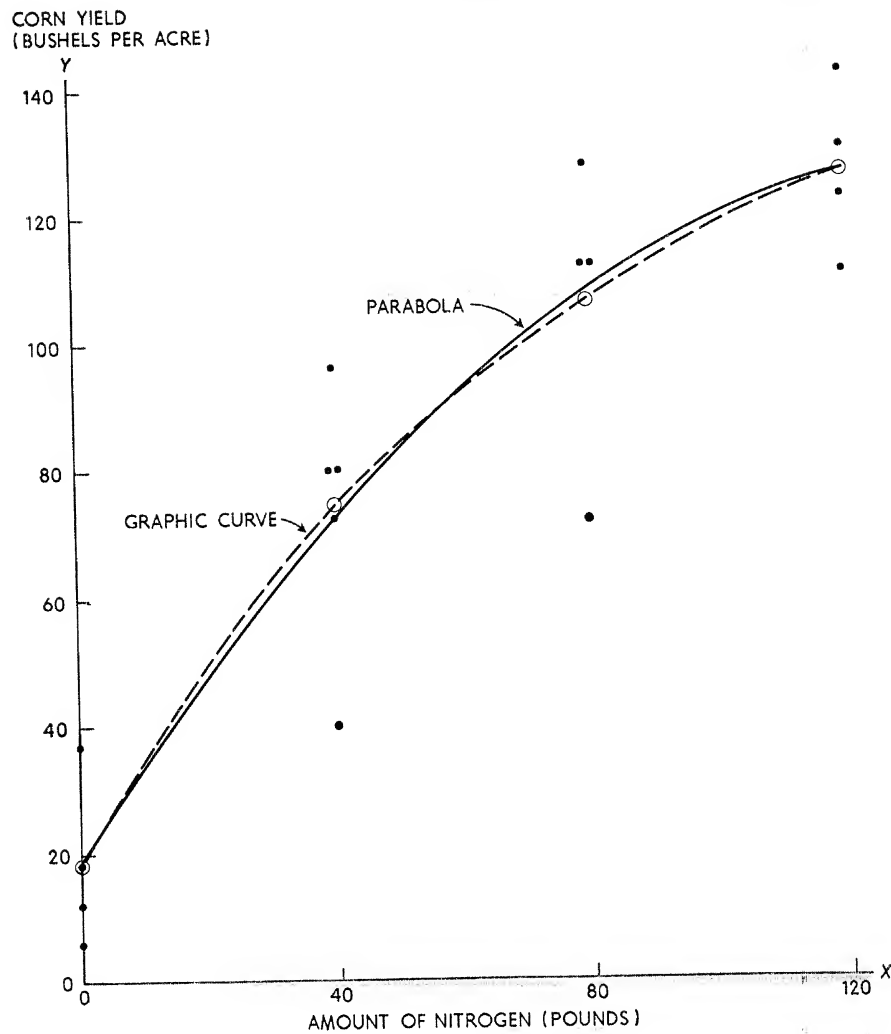
along the X axis, the graphic curve would go close to the group averages, although not necessarily passing through all of them.

If the relationship is really curvilinear, a hand-drawn curve is likely to be a better fit than a straight line fitted by least squares, however impressive the mathematical formula and computer used. The analyst should always plot his data, check for curvilinearity, and consider whether the relationship is logically curvilinear rather than automatically using some straight-line computer program.

Multiple Regression. The graphic method is also helpful in determining net curvilinear relations in multiple regression when the appropriate mathematical equation is not known. The same method of successive elimination may be employed as described in Chapter 23, except that curves are drawn instead of straight lines. As a short cut, the dependent variable may be first plotted against one of the independent variables and a graphic regression curve drawn; then the vertical residuals from this curve ($z = Y - Y_c$) are plotted above and below the zero line, with the second independent variable as abscissa. A second

Chart 24-1

NITROGEN FERTILIZER AND CORN YIELDS
SIXTEEN FIELDS



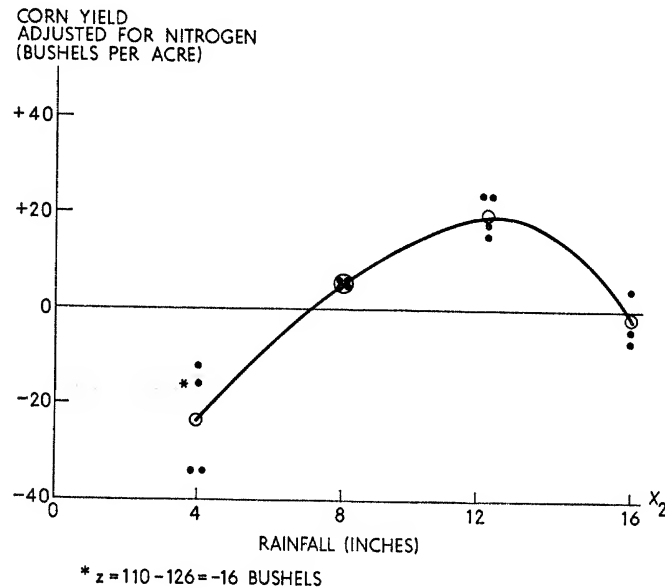
curve is drawn, and the residuals from this curve are in turn plotted against a third independent variable (if any) or else are laid off around the first regression curve. The curve is redrawn, and this process is refined by transferring the residuals back and forth until no further improvement occurs in the net regression curves. Few approximations will be required if the independent variables are not correlated with each other. An alternative method is first to compute a multiple linear regression and then plot the residuals against each of the independent

variables in turn and draw freehand curves to adjust the preliminary linear relationship for curvature.¹

In the corn yield experiment (Chart 24-1), nitrogen accounted for a good deal of the increase in yields, but not all, since the dots deviated considerably from the regression curve. What other influences were at work? Suppose rainfall during the growing season varied from 4 to 16 inches for the 16 fields. We can plot the residuals (z) from Chart 24-1 (i.e., the variation in yield *not* explained by fertilizer) against rainfall

Chart 24-2

RAINFALL AND CORN YIELDS ADJUSTED FOR CHANGES IN NITROGEN
SIXTEEN FIELDS



in Chart 24-2 to see if this variation can be explained by rainfall. (Comparative data are not listed here.) By drawing group averages and a freehand regression curve in Chart 24-2, we find that rainfall up to about 12 inches stimulates yields, but heavier rainfall depresses yields, quite apart from most of the effect of nitrogen. Thus, the first field receiving 120 pounds of nitrogen yielded only 110 bushels (Table 24-1) compared with its expected yield of 126 bushels (Chart 24-1). This deficit of 16 bushels, however, is partly explained in Chart 24-2, since the field received only 4 inches of rain (see asterisk); thus one would expect a deficit of 24 bushels, so that the remaining unexplained

¹ See M. Ezekiel and K. A. Fox, *Methods of Correlation and Regression Analysis*, 3d ed. (New York: John Wiley, 1959), chaps. 14 to 16, for a detailed discussion of multiple curvilinear regression.

variation in yield is only +8 bushels. Further approximations would refine these results.

We can forecast corn yields by adding the values on the two regression curves for any combination of fertilizer and rainfall represented by our experiment. Thus, for fields with 120 pounds of fertilizer and 4 inches of rain, we would expect a yield of 126 bushels (Chart 24-1) minus 24 bushels (Chart 24-2) or 102 bushels, on the average.

This experiment also illustrates the "regression model" (Chapter 22) in which regression results are only valid for the selected values of X_1 and X_2 , and the coefficient of correlation is of doubtful significance because of the arbitrary limits placed on these values.

Fitting Mathematical Curves

Graphic methods have a certain flexibility in that the curve can be drawn to fit the data as closely as desired. Mathematical methods, on the other hand, have the advantage of fitting a curve (or surface) that can be described by an equation. This makes it somewhat easier to summarize the relationships, evaluate the results, and predict new observations. However, the degree of success in fitting a mathematical relationship depends upon how carefully the functional form of the equation is picked. There are polynomials, logarithmic functions, and many others. We shall next examine the first two of these functions as used in simple regression.

The Parabola. The simplest curve is the parabola of the form $Y_c = a + bX + cX^2$. In this equation, a is the height of the curve at the Y axis, b is the slope of the curve at this point, and c determines the direction and degree of curvature.

To fit a parabola,² we can treat X^2 as if it were a new variable X_2 . Then, if we call the original variable X_1 , and change the constants b

² Alternatively, if we use x and y to represent deviations of X and Y from their means, we can solve the following two normal equations to determine the values of b and c in the original equation:

$$\begin{aligned}\Sigma xy &= b\Sigma x + c\Sigma x^2 \\ \Sigma x^2y &= b\Sigma x^2 + c\Sigma x^3\end{aligned}$$

The constant term a can then be calculated from the formula:

$$a = \bar{Y} - b\bar{X} - \frac{c\Sigma X^2}{n}$$

Here, \bar{X} , \bar{Y} , Σx^2 , and Σxy have already been defined and

$$\begin{aligned}\Sigma x^3 &= \Sigma X^3 - \bar{X}\Sigma X^2 \\ \Sigma x^4 &= \Sigma X^4 - \frac{(\Sigma X^2)^2}{n} \\ \Sigma x^2y &= \Sigma X^2Y - \bar{Y}\Sigma X^2\end{aligned}$$

and c to b_1 and b_2 , respectively, the equation for the parabola becomes $Y_c = a + b_1X_1 + b_2X_2$. This is identical with the equation for multiple regression (Chapter 23), so we can use the same techniques to find a , b_1 , and b_2 . In particular we solve the normal equations:

$$\begin{aligned}\Sigma x_1y &= b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 \\ \Sigma x_2y &= b_1\Sigma x_1x_2 + b_2\Sigma x_2^2\end{aligned}$$

and then $a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$.

Here, the variables are expressed as deviations from their means: $y = Y - \bar{Y}$, $x_1 = X_1 - \bar{X}_1$, and $x_2 = X_2 - \bar{X}_2$.

A parabola has been fitted to the corn-yield data in Table 24-1 with the following result:

$$Y_c = 18.6 + 1.565X - 0.005625X^2$$

The parabola is plotted in Chart 24-1. The curve does not pass precisely through the means of the four arrays, though it comes close to doing this. The parabola and graphic curves fit the data about equally well. The parabola is more objective, while the graphic curve is more flexible in being able to approximate types of functions that cannot be represented by simple mathematical formulas.

The method used here for fitting the parabola is generally applicable to higher order polynomials. For example, the cubic polynomial is $Y_c = a + bX + cX^2 + dX^3$. By defining $X = X_1$, $X^2 = X_2$, and $X^3 = X_3$, we can fit the cubic by using the normal equations for multiple regression with three independent variables.

Use of Logarithms. If the relationship appears curvilinear when plotted on an arithmetic grid, the data can be replotted on semilogarithmic graph paper (with either variable on the log scale) or on a double-logarithmic graph. Then, if the data follow approximately a *straight* line on any of these charts, the line can either be drawn graphically with a ruler or fitted by least squares.

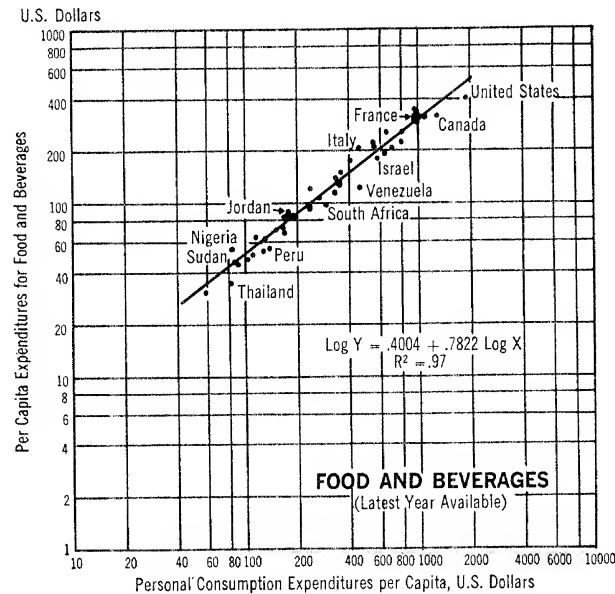
In the least squares method, the *logarithms* of the appropriate variable(s) are used in place of the original values, and a straight line is fitted just as described in Chapter 22. Thus, if the relationship is linear when plotted on semilogarithmic paper (with Y on the log scale), the equation of the regression line is $\log Y_c = a + bX$. The method of fitting this equation in trend analysis was illustrated on pages 490-495. Conversely, a straight line on semilogarithmic paper with X on the log scale has the form $Y_c = a + b \log X$. Finally, if the relationship is linear when plotted on double-logarithmic paper, the equation is $\log Y_c = a + b \log X$. This equation is a reasonable one to use if Y tends to change by a constant *percent* for each 1 percent change in X over all X values. In Chart 24-3, for example, expenditures for food and bev-

erages are plotted against personal consumption expenditures for selected countries on a double-logarithmic scale. A straight line describes the relationship fairly closely.

Other Transformations. The use of logarithms is a special case of the more general technique of *transformation* of variables to achieve

Chart 24-3

EXPENDITURES FOR FOOD AND BEVERAGES AND
TOTAL PERSONAL CONSUMPTION FOR SELECTED
COUNTRIES



SOURCE: Mary K. Baird, *International Consumer Expenditure Patterns* (Menlo Park, California: Long Range Planning Service, Stanford Research Institute, 1963), p. 10, with permission.

straight-line relationships. In that instance, the variable X (or Y) was transformed into $\log X$ (or $\log Y$), and a linear regression equation was calculated, using the transformed values in place of the original data.

If the logarithmic relationship is not linear, we can transform the variables into another function in order to get a linear fit. Some common transformations include the use of the square root, the reciprocal, e^x , and combinations of these. Many computer programs incorporate these transformations automatically in the computation of the regression equations.³ The question of which transformation to use in a

³ See *BMD Biomedical Computer Programs*, pp. 15 to 21, for a list of more than 20 transformations or "transgenerations" available in those programs. (Health Services Computing Facility, School of Medicine, University of California, Los Angeles, Jan. 1, 1964.)

specific situation is one of judgment and experience. The analyst should select functions that make sense logically and he should then try several until he finds one that produces a satisfactory linear fit.

The use of transformations is even more important in multiple curvilinear regression than in simple curvilinear correlation, since with many variables the calculations and their interpretation are much simpler using the first power of transformations of X , rather than becoming involved with higher powers of X .

Transformations and Homoscedasticity. Generally, for valid conclusions from regression analysis, the data must be uniformly scattered about the regression line or plane. When this assumption of homoscedasticity is not satisfied, a transformation of the data may serve to produce a more even dispersion. For example, if the scatter about the regression line tends to be a constant *percent* of the independent variable X , then the use of $\log Y$ will make the absolute deviations about the line more uniform.

Standard Error of Estimate

Just as in linear regression, the *standard error of estimate* is used to measure how closely the curvilinear equation fits the data. The standard error of estimate for any number of variables is

$$S_{YX} \text{ or } S_{Y \cdot 12 \dots} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - k}}$$

Here $(Y - Y_c) = z$ is the deviation of the dependent variable from its computed value (determined either mathematically or graphically); the term n is the number of observations; and k is the number of constants in the regression equation. If a graphic curve is used, k is estimated as the number of constants that would occur in a mathematical curve of the same general shape.

Index of Correlation

The index of correlation (or its square, the index of determination) is used in curvilinear correlation as a relative measure of association varying from 0 (no correlation) to 1 (perfect correlation). No sign is used. The index of correlation may be found in the same way as the coefficient of correlation in linear correlation by computing $\sqrt{1 - S_{YX}^2/S_Y^2}$, where S_{YX} is the standard error of estimate and S_Y is the standard deviation of the dependent variable. The formula applies in both simple and multiple curvilinear correlation.

Statistical Inference

The techniques we have discussed for mathematical curvilinear functions all involve conversion of the relationship to one of linear regression, either by transformations or by defining new variables (such as letting X^2 be X_2). Once we have done this, the methods of making statistical inferences for the linear regression model are also applicable to the transformed data. For example, if we were to fit the function $Y_c = a + b \log X$, the calculation of the standard error of b and tests of hypothesis and confidence intervals for b could be determined in the normal manner. We simply substitute $\log X$ in place of X in the appropriate formulas.

When to Use Curvilinear Methods

Curvilinear measures of regression should be used whenever (1) the *rationale* of the situation calls for a curved relationship and (2) the curve actually fits the data better than a straight line, as measured by the standard error of estimate. Thus, in measuring the effect of nitrogen on corn yield, it is logical to expect diminishing returns since successive increments of nitrogen should produce smaller and smaller increases in yield up to a maximum, after which yield should drop with an excess of fertilizer. Hence, a parabola or freehand curve concave downward is *a priori* superior to a straight line.

Second, as a measure of goodness of fit for the corn-yield experiment (Table 24-1), the standard error of estimate around the parabola is

$$s_{YX} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - k}} = \sqrt{\frac{4,521}{16 - 3}} = 18.6 \text{ bushels per acre}$$

A straight line (not shown) was also fitted by least squares to the same 16 observations. Its equation is $Y_o = 27.6 + 0.89X$, and its standard error is

$$s_{YX} = \sqrt{\frac{\Sigma(Y - Y_o)^2}{n - k}} = \sqrt{\frac{5,817}{16 - 2}} = 20.4 \text{ bushels per acre}$$

It appears that the parabola *does* give more accurate estimates than the straight line, since the average scatter is smaller for the curve even after allowing for the increase in k , the number of constants in the equation.

In other situations the same *percent* increase in Y may logically

follow a 1 percent increase in X , as noted above. Here, it is rational to fit a straight line to the logarithms of the data. Other transformations should also be justified on rational grounds. Finally, in predicting production ratings from test scores, we knew of no *a priori* reason why the regression should be curvilinear, and the actual plot followed a straight line, so that linear analysis was justified.

CORRELATION OF TIME SERIES

The correlation of time series presents no new computational problems. The analysis of two series ordered in time may be carried out in a manner similar to that illustrated in the previous chapters. Problems of interpretation do arise, however, and there are some "booby traps" for the novice.

In the first place, much of the observed correlation between two economic time series may be due to the fact that both variables have strong upward trends. Any two linear trends will be perfectly correlated with one another, whether the series has any real connection or not. In appraising a high coefficient of determination obtained between total meat consumption and disposable income over a 40-year period, we should recognize that population growth is the most important component in the dependent variable and that it also accounts for about half of the increase in disposable income (if the latter is expressed in constant prices). This is a cheap and unenlightening victory; other things being equal, it is perfectly obvious that two people will consume twice as much meat as one. If *economic* relationships are important, these relationships should be investigated on a *per capita* basis. Further investigation may indicate that this relationship cannot be established satisfactorily by means of simple regression analysis but that it calls for multiple regression analysis.

In other cases, there may be trends in time series due to factors other than population growth or general growth of the economy. It must be decided whether interest is best served by (1) explaining the trend and ignoring the year-to-year fluctuations, as in Chapter 19, (2) eliminating the trend and explaining the year-to-year fluctuations, or (3) attempting to explain both simultaneously.

Methods of Correlating Time Series

There are four ways to correlate time series. The first two of these will be illustrated in the correlation of photographic equipment sales and disposable personal income listed in Table 24-2. The following discussion applies to annual data; monthly data should be adjusted for

calendar and seasonal variation before being correlated. Dollar value series may be correlated without price adjustment, as in our example, if it is wished to compare the combined effect of changes in price and physical volume. However, if it is desired to bring out the relationship in physical volume changes unobscured by price fluctuations, the dollar series should be deflated by dividing through by an appropriate price index. Unfortunately, however, each deflated series is affected by an unknown error in the price index itself. The four methods of correlating time series are as follows:

1. *Correlate the actual annual data* to show the combined effect of secular trend and cyclical and irregular fluctuations. This method may be quite adequate for forecasting, particularly over the longer run. The pitfall here is that any two series that have nonhorizontal trends or that are affected by the general business cycle will *appear* to be correlated whether or not there is any real connection. The meat consumption example was a case in point. The remedy is to (1) choose only series that have a close logical relationship; (2) supplement this method with one of the following, in which trend is eliminated; and (3) avoid the coefficient of correlation or determination, which is spuriously high.

2. *Correlate first differences*, such as the percent changes from a year ago listed in Table 24-2. The use of these percents will eliminate all trend except that in a single year and will avoid the errors involved in fitting a trend curve (method 3). This method is useful chiefly in short-term forecasting. Either the *relative* first differences (percent changes) or *absolute* first differences (amounts of change) may be correlated. The amounts of change are obtained by *subtracting* each year's values from the next. It is usually better to correlate relative rather than absolute first differences, since percents tend to have a more uniform dispersion over a period of time than do the absolute amounts. For example, the year-to-year changes in the dollar volume of photographic equipment sales tend to become larger in later years simply because the sales volume itself is so much greater. The later values thus have a disproportionate influence in determining the various measures of correlation, if absolute values are correlated.

3. *Correlate percents of trend*, that is, cyclical-irregular relatives. These values are shown in Chart 19-6 for Sears, Roebuck sales. Similar deviations could be determined for sales of photographic equipment and disposable income. The results bring out the *cyclical* and other short-term relationships between the two series. This method, therefore, is useful for anticipating the effect of short-term business cycle changes. The trend line is a more stable base for computing percents than is the

previous year's level, so the scatter of percents tends to be less erratic than in method 2. However, in the long run the projections obtained in method 3 are increasingly sensitive to errors in extrapolating the trend curve itself.

4. *Apply multiple regression analysis*, with time as a separate independent variable. Thus, we could correlate photographic equipment sale with both disposable income and years. The regression coefficients then give the separate influence of income and trend (years) on sales, unless the independent variables are too closely correlated with each other to permit these influences to be segregated. If this is done, the equation form must be consistent with the secular trend function. Since sales of photographic equipment follow a logarithmic straight line trend of the form $\log Y_c = a + bX$, the *logarithm* of sales should be used in a linear multiple regression if time is used as one of the independent variables. Otherwise, the results will be distorted. If trends are present, logarithms are often used for *all* variables except time, in order to make the scatter of the residuals more uniform than would be the case if absolute values were used. By this device all correlation measures may become more meaningful.

Correlating Actual Data. Suppose we are engaged in long-range planning for Eastman Kodak Company and wish to establish a quantitative basis for projecting the company's future sales. Total U.S. expenditures for photographic equipment should logically be related to disposable personal income. Increases in disposable income reflect the growth both in population and in affluence (i.e., in per capita income). Each of these factors should stimulate sales of photographic equipment. Therefore, we will correlate photographic equipment sales with disposable personal income. The sales forecast for Eastman Kodak can be determined from the industry sales projection by estimating the Eastman Kodak percent share of the market.

Sales of photographic equipment⁴ and disposable personal income for the years 1948–1963 are shown in Table 24–2. This 16-year period was selected to exclude World War II and the immediate postwar readjustment years, as well as the two latest years, which are held out as a check on the forecast. The regression equation for 1948–1963 will be used to predict sales for 1964 and 1965. We will then check the forecasts against actual sales in these years.

The first step is to plot the data on a scatter diagram. Either an

⁴ Photographic equipment sales represent total sales of Polaroid, Eastman Kodak, and Bell and Howell, from *Moody's Industrial Manual*. This comprises a very substantial part, but not all, of total industry sales.

Table 24-2

PHOTOGRAPHIC EQUIPMENT SALES AND DISPOSABLE PERSONAL INCOME
1948-1963 AND FORECAST YEARS 1964-1965

Year	Disposable Income, Billions of Dollars X	Sales of Photographic Equipment, Billions of Dollars Y	Percent Change from Previous Year	
			X	Y
1948	189.1	0.457		
1949	188.6	0.418	-0.3	-8.5
1950	206.9	0.488	9.7	16.7
1951	226.6	0.579	9.5	18.6
1952	238.3	0.625	5.2	7.9
1953	252.6	0.704	6.0	12.6
1954	257.4	0.713	1.9	1.3
1955	275.3	0.800	7.0	12.2
1956	293.2	0.867	6.5	8.4
1957	308.5	0.929	5.2	7.2
1958	318.8	0.985	3.3	6.0
1959	337.3	1.109	5.8	12.6
1960	350.0	1.158	3.8	4.4
1961	364.4	1.204	4.1	4.0
1962	385.3	1.308	5.7	8.6
1963	403.8	1.389	4.8	6.2
Average	278.26	.8583	5.21	7.88
Future Years (Actual)				
1964	435.8	1.548	7.9	11.4
1965	465.3	1.853	6.8	19.7

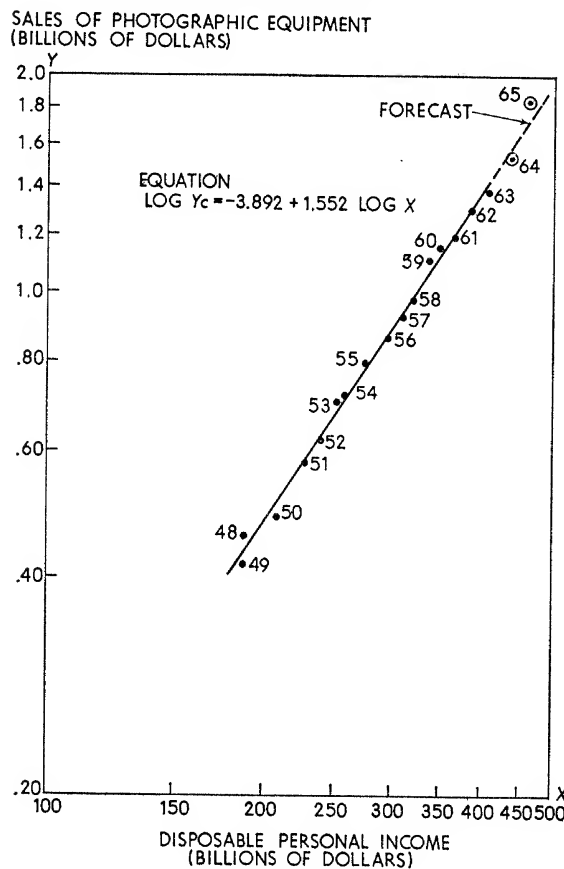
Source: *Business Statistics* (1965) and *Survey of Current Business*; *Moody's Industrial Manual* (1965) and company annual reports.

arithmetic scale or a logarithmic scale can be used. In this case, the double logarithmic scale was selected both because the scatter of dots appeared more linear on this scale *empirically*, as shown in Chart 24-4, and because relative (percent) changes should *logically* have a more linear relation than absolute amounts of change. From the logarithms of the data in Table 24-2, the regression line is computed by least squares as $\log Y_e = -3.892 + 1.552 \log X$. This line is plotted in Chart 24-4. (The natural values are plotted, not the logs.)

For more extended analysis, we could compute the standard error of estimate and confidence intervals for both the regression line and an individual forecast, as we did for the production rating example in Chapter 22. The confidence interval for the regression line would apply

Chart 24-4

SALES OF PHOTOGRAPHIC EQUIPMENT AND
DISPOSABLE PERSONAL INCOME 1948-1963,
WITH FORECASTS FOR 1964 AND 1965



SOURCE: Table 24-2.

if we were forecasting the general level or trend of sales over a number of future years, while the confidence interval for an individual forecast would apply if we were predicting sales for a particular year. We will not repeat this procedure here, as the regression line will suffice to illustrate the peculiarities of correlating time series.

Most measures of correlation and regression are theoretically correct only if the residuals ($Y - Y_c$) are randomly distributed, with uniform dispersion, around each section of the regression line, as described in Chapter 22. This is not true of time series. First, the presence of an extreme high or low value (occasioned, say, by a war scare or strike) influences the regression line in proportion to the *square* of its deviation and so distorts the line.

Second, the absolute residuals tend to get bigger as the industry grows over the years. The use of logarithms to discount this tendency is illustrated in the present example.

Third, since most time series move in cycles rather than in purely random fashion, there are likely to be runs of several successive positive residuals or several negative residuals in a row. That is, each year's value is related to that of the adjoining year rather than being independent of it. This is called "autocorrelation." If autocorrelation exists in the residuals, the standard error of estimate will understate the amount of error likely to be encountered in making forecasts for one or two years ahead. Essentially, autocorrelated series give us less information per observation than do completely random ones. The closer together in time we take our observations, the greater will be the autocorrelation between them. Hence, seasonally adjusted monthly data will exhibit a higher degree of autocorrelation than annual data.

Tests are available for appraising the extent of autocorrelation in the residuals from a time series analysis, but these tests will not be described here.⁵ If the degree of autocorrelation is greater than could be attributed to chance, the usual standard error formulas are inapplicable. There is some autocorrelation evident in Chart 24-4 since there are several sequences or runs of years above and below the regression line. For example, the years 1953-1955 are all above and the years 1956-1958 are all below the regression line.

Correlating First Differences. With most economic time series, positive autocorrelation in residuals is found when the original values of the variables are correlated. Positive autocorrelation can usually be reduced by using first differences, as in the last two columns of Table 24-2. If the regression equation is calculated in terms of first differences and the residuals from this equation are not significantly autocorrelated, then the standard errors of regression coefficients and the standard error of estimate are regarded as applicable and valid for the span of years covered. Use of this equation for forecasting in subsequent years still depends upon the study of future trends that would affect the relationship. This topic will be discussed later.

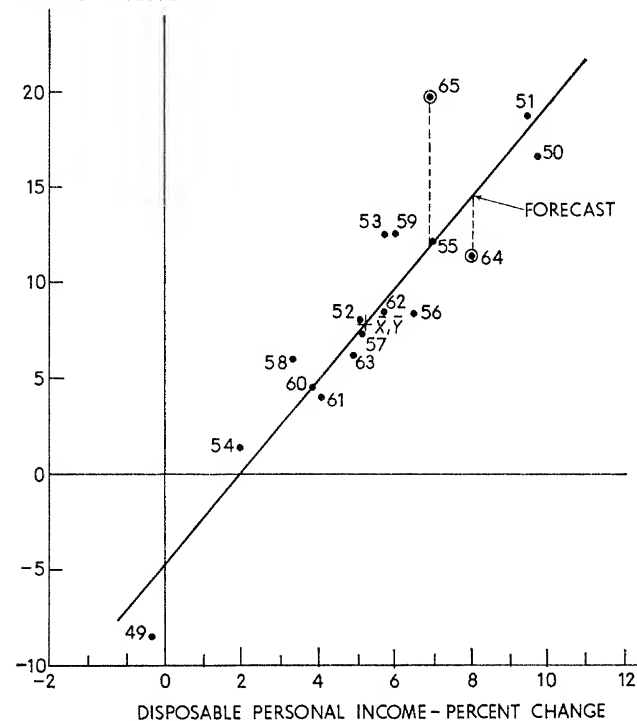
The relative first differences, or percent changes from a year ago, are plotted in Chart 24-5 for sales of photographic equipment and disposable personal income. A regression line has been fitted to these points by the method of least squares. The residuals in Chart 24-5 appear to be more randomly distributed than those in Chart 24-4, although there is

⁵ The principal tests are the "coefficient of autocorrelation" and "von Neumann's ratio." For details, see M. Ezekiel and K. A. Fox, *Methods of Correlation and Regression Analysis*, 3d ed. (New York: John Wiley, 1959), pp. 334-40.

Chart 24-5

PERCENT CHANGES FROM PREVIOUS YEAR IN SALES OF
PHOTOGRAPHIC EQUIPMENT AND DISPOSABLE INCOME
1949-1963, WITH FORECASTS FOR 1964 AND 1965

SALES OF
PHOTOGRAPHIC EQUIPMENT
PERCENT CHANGE



SOURCE: Table 24-2.

still considerable autocorrelation. For example, while the years 1956 through 1958 no longer fall on the same side of the line, the years 1951-1955 are all above the line. The various standard errors computed for these percents, nevertheless, should be slightly more valid than those computed for the original values in Chart 24-4. This does not mean, of course, that a *forecast* based on first differences is necessarily more accurate than one based on original data.

Is the Correlation of Aggregates Valid for Forecasting?

The data used in this analysis were totals for the entire United States over the period 1948-1963. Each variable—net sales and disposable personal income—is essentially a population total, so there is no room for sampling errors in the variables, although there may be some errors of measurement.

Do the data form a sample in any sense or do they simply describe a condition of the population in a particular time period? It may be assumed that the dependent variable, sales of photographic equipment, is not perfectly correlated with disposable income, but is subject to a large number of more or less random disturbances in the economy. These forces may be too small and too numerous to list and measure separately; they are not predictable in advance, so that their net effect in the year just ahead is just as likely to raise the dependent variable above its average relationship to disposable income as it is to lower the dependent variable. Hence, the values of the dependent variable will include a systematic component, related to disposable income, and also a random component. The systematic component is estimated by fitting a regression line; the random component will be reflected in the residual variation of the dependent variable around this line.

Thus, the population with respect to which the 1948–1963 observations on disposable income and photographic equipment sales have sampling significance is a rather peculiar one—a population in which the same set of values of disposable income is repeated time after time but in which the random economic disturbances will give rise to different observed values of sales for any given value of disposable income in successive samples.

How can the 1948–1963 regression equation be used in later years? If there is no reason for the random economic disturbances to increase or decrease in magnitude, we may tentatively assume that the standard error of estimate computed for 1948–1963 will continue to apply. But will the regression equation hold good in these later years? The equation is $\log Y_e = -3.892 + 1.552 \log X$. For this relationship to remain valid, percentage increases in sales of photographic equipment will have to continue to account for about the same fraction of the percentage increases in consumers' disposable income as it has in the past. To test this assumption, a more detailed analysis would be necessary, including studies of changes in products, in advertising, in consumer preferences, and in general economic conditions. If there is evidence that photographic equipment will constitute a greater share of the consumer's dollar purchases in the future, then the regression equation will have to be modified accordingly.

Attitudes toward extrapolation of regression curves differ greatly. Many writers insist that a regression function must not be applied beyond the range of the data on which it is based. On the other hand, estimates needed for practical purposes are sometimes obtained by reckless extrapolation of regression functions. Both extremes should be avoided. One of the major purposes of regression analysis is to provide

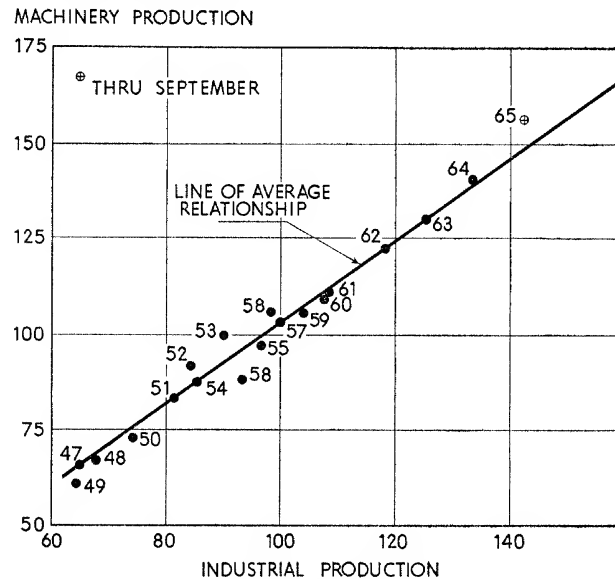
the basis for estimates, and these sometimes involve extrapolation. At the same time, the analyst should be aware of difficulties associated with extrapolation and should support his statistical analysis with a good logical justification for any extension of a regression beyond the limits of the data on which it is based.

Forecasting Sales

Industrial output or company sales are often forecast by correlation with some basic measure of the economy—such as gross national prod-

Chart 24-6

RELATIONSHIP OF MACHINERY PRODUCTION TO
INDUSTRIAL PRODUCTION (1957-1959 = 100)



SOURCE: Standard & Poor's, *Industry Surveys, Industrial Machinery* (October 28, 1965).

uct, disposable personal income, or industrial production—for which relatively reliable projections are available. Chart 24-6 shows the relation of machinery production to total industrial production—one of many similar charts presented in Standard & Poor's *Industry Surveys*. Forecasts of industrial production and other basic economic indicators, made by numerous agencies for periods up to 15 years in the future, are reported in *Predicasts*, published quarterly by Economic Index & Surveys of Cleveland, Ohio.

In the photographic equipment example, assume that we have accurate projections of disposable income—\$435.8 billion for 1964 and \$465.3 billion for 1965. We can then forecast sales for these years from

the regression line in Chart 24-4. Later on, in 1966, we can check these forecasts against actual sales. These are circled on the chart. The results (in billions of dollars) are shown in the table.

Year	Forecast Sales	Actual Sales	Forecast Error, Percent
1964	1.602	1.548	3.5
1965	1.773	1.853	-4.3

We can also forecast sales for these two years from the projected *percent* changes in disposable income shown in Chart 24-5. These forecasts compare with actual results as shown in the table.

Year	Increase in Disposable Income, Percent	Forecast Increase in Sales, Percent	Resulting Sales Forecast, Billions	Actual Sales, Billions	Forecast Error, Percent
1964	7.9	14.4	1.589	1.548	2.6
1965	6.8	11.1	1.765	1.853	-4.7

Of course, this analysis does not include errors in projecting disposable income itself; such errors would either increase or decrease the error in the photographic equipment sales forecast.

These forecasts illustrate the basic premise upon which projections are made: Extrapolation or interpolation from past data is valid only if the basic underlying relationship remains the same. For 1964, this is true, and the forecasts for that year were reasonably accurate. For 1965, however, the forecasts were less accurate. This resulted, in part, from the introduction of new products in 1965—Polaroid introduced low-priced cameras for the first time, increasing sales nearly 50 percent, and Kodak introduced the Instamatic camera. Extrapolations based upon past data need to be adjusted for such innovations in products as well as for changes in management policy and shifts in consumer behavior as they occur.

Furthermore, when projecting company sales, the analysis should be carried out separately for individual lines of merchandise and for different territories. It may then be possible to pinpoint with considerable assurance certain sources of demand that will behave about the same in the late 1960's as they did in the 1950's and early 1960's and other sources of demand that may change drastically. Correlation analysis may

be our basic tool in determining these more detailed relationships, and our forecasts of total sales could very well be the sum of forecasts of its individual components based on these several regression equations.

SUMMARY

Often a straight line does not adequately represent the relationship between two variables. In such cases, a freehand or mathematical curve may better fit the data.

In fitting a curvilinear function graphically, the data and several group averages are first plotted. A smooth curve is then drawn through the group averages or as close to them as possible. If there is more than one independent variable, the vertical deviations from the first regression curve may be plotted against the second independent variable; another regression line is drawn, and the deviations from this curve are drawn against a third variable, or against the first regression curve, which is then redrawn, and so on until the curves stabilize. This is the "method of successive elimination."

Many mathematical functions also may be used to express a curvilinear relationship between two or more variables. The most common are the *parabola* and the *logarithmic straight line*.

A parabola is a curve of the form $Y_c = a + bX + cX^2$. It may be fitted by treating the X^2 term as a new variable X_2 and then solving the normal equations for multiple regression, using the redefined variables.

To fit a logarithmic straight line, the data may be plotted on semilog or double-log graph paper and a straight line drawn graphically. Alternatively, logarithms may be used in place of any or all of the variables in the calculations of the least squares regression line.

The use of logarithms in regression equations is an example of the *transformation* of variables. Other transformations, such as the use of square roots or reciprocals, may also be used in regression analysis to produce a good curvilinear fit.

As in linear regression, the *standard error of estimate* measures the average error of the regression curve in providing estimates of Y from given values of the independent variables. It is the standard deviation of the residuals ($z = Y - Y_c$) adjusted for the number of constants in the regression equation.

The *index of correlation* measures the proportion of the variation in the dependent variable that is accounted for by the independent variables. It is equivalent to the coefficient of correlation in linear correlation.

The methods applicable in estimating curvilinear relationships may be used for any number of variables. Statistical inferences may also be drawn from the curvilinear regression of sample data by the same methods as in linear regression, provided the variables have been transformed into linear form.

Curvilinear methods of regression should be used whenever (1) the logic of the relationship supports a particular type of curve and (2) the standard error of estimates is smaller for this curve than for a straight line.

Regression techniques are applicable to time series, provided there is a rational hypothesis supporting the relationship. There are four methods of correlating time series, of which the first two are illustrated in this chapter.

1. *Correlate the actual annual data* (or deseasonalized monthly data) to show the combined effects of secular trend and cyclical and irregular fluctuations.
2. *Correlate relative or absolute first differences* (percents or amounts of change from year to year) to partially eliminate trend.
3. *Correlate percents of trend*, using secular trend values as a base. Methods 2 and 3 show the relationships of cyclical and other short-term fluctuations.
4. *Apply multiple regression analysis* with time as one independent variable. Logarithms may be used for all variables except time to achieve a more uniform scatter of residuals.

Photographic equipment sales are correlated with disposable personal income for 1948–1963, and the regression is used to forecast 1964 and 1965 sales. The results are then checked against the actual sales for these years. Plotting the original data in Chart 24–4, we find a close linear relationship. However, there is danger that the residuals around the line may be autocorrelated (i.e., successive years' values may be alike), so the standard error formulas may be inapplicable. In order to reduce autocorrelation and eliminate trend, which produces a spuriously high correlation, we plot the year-to-year percent changes in Chart 24–5. The relative scatter here is wider, but the various standard error formulas are more valid than in correlating original data.

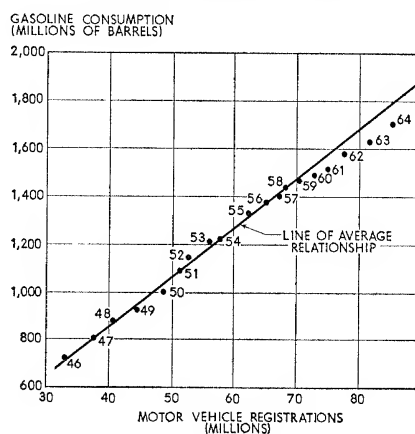
To determine whether regression relationships will apply in the future, one must make a careful study of management policy, consumer preferences, and general economic trends. Extrapolation of regression

curves is dangerous, but it is nevertheless necessary and widely used in forward planning. The forecasts of photographic equipment sales based on the regression with disposable income proved to be fairly accurate for 1964 but less so for 1965. For management planning purposes, more elaborate correlation and trend analysis is needed, as well as a careful appraisal of future policy and qualitative economic factors. This analysis should be applied to individual products, territories, and departments of the business.

PROBLEMS

1. As an oil company economist you wish to forecast U.S. gasoline consumption, in barrels, for each of the next five years by correlation with some basic economic factors for which forecasts are available. One such factor—motor vehicle registrations—is illustrated below. Give two other factors that you might logically choose to correlate with gasoline consumption as a basis for prediction. Support your choice.
2. The regression line in the chart was fitted to data through the late 1950's and extended to provide a forecast of gasoline demand in future years. This forecast, however, proved to be too high, as shown by the actual data for 1960–1964.

RELATIONSHIP OF DOMESTIC CONSUMPTION OF GASOLINE TO MOTOR VEHICLE REGISTRATIONS



SOURCE: Standard & Poor's, *Industry Reports, Oil* (November 25, 1965).

- a) What economic developments in these later years might have caused this shift in the regression line?
- b) What economic assumptions would you have to make in order to justify fitting a new regression line to the 1955–1964 data to forecast 1970 gasoline demand, based on an available estimate of motor vehicle registrations in that year?

3. As an experiment, Sears, Roebuck sales were correlated with disposable personal income for the years 1947–1956 by each of the four methods described on pages 640–646, with the following results:

Factors Correlated	Standard Error of Estimate	Coefficient of Determination
(1) Actual annual data	\$110 million	0.93
(2) Relative first differences (Percent changes)	6.74 percent	0.40
(3) Percent of log straight line trends	3.25 percent	0.43
(4) Multiple regression between actual sales, disposable income, and number of stores	\$115 million	0.93

Sears sales in 1956 were \$3,601 million, while the 1957 trend value was about \$3,700 million.

- a) In the light of this information, which of these four methods would have been preferable for use in forecasting 1957 sales, based on an available estimate of 1957 disposable income? Why?
- b) Is disposable income satisfactory as a predictor of short-run changes in Sears sales? Explain your answer.
4. Retail sales in recent years were as follows:

RETAIL SALES IN THE UNITED STATES, 1951–1964
(Billions of Dollars)

Year	Durable Goods	Nondurable Goods
1951	54.5	102.1
1952	55.3	107.1
1953	60.4	108.7
1954	58.1	111.0
1955	67.0	116.9
1956	65.8	123.9
1957	68.5	131.5
1958	63.4	136.9
1959	71.7	143.7
1960	70.7	148.8
1961	67.3	151.5
1962	74.9	160.4
1963	80.1	166.3
1964	85.1	176.5

SOURCE: U.S. Department of Commerce, *Survey of Current Business*.

- a) Plot retail sales of either durable goods or nondurable goods, as assigned, against disposable personal income (Table 24–2) for 1951–1964 on an arithmetic scatter diagram.
- b) Fit a regression line by the graphic method or by least squares, as assigned, and draw a band one standard error of estimate above and below it, as a rough 67 percent confidence interval. Describe the relationship in these years and the probable reason for the deviation of points from the line.
- c) Forecast 1965 retail sales of durable or nondurable goods and give the

- $\pm 1S_{YX}$ limits, based on estimated 1965 disposable income of \$465.3 billion.
- d) Compare this forecast with actual 1965 sales of \$94.7 billion for durables or \$189.1 billion for nondurables (based on 1964–1965 percent increases in the new series applied to 1964 sales above). Explain the probable reason for your error of forecast.
5. a) How could you determine whether the regression between test scores and production ratings in Table 22–5 is significantly curvilinear?
- b) Since the formula for a straight line is merely a special case of that of a parabola in which $c = 0$, the parabola would seem to fit almost any set of data better than the less flexible straight line. Can you infer, then, that nearly all regressions are significantly curvilinear? Explain.
6. The Value Line Investment Survey computes a multiple regression equation for each common stock showing the typical relationship between its price (X_1), earnings per share (X_2), and dividends per share (X_3) in past years. The following equation was reported for Boeing Airplane Company:
- Log normal average value next 12 months
 $= 1.355 + 0.440 \log (.22 \times \text{earnings} + 1.00 \times \text{dividends})$
- a) Explain the meaning of this equation and its use for an investor.
- b) What type of linear transformation does this equation illustrate?
- c) What other measures or qualifications would be desirable in this survey to aid the investor in appraising the reliability of the equation?
7. You are an analyst interested in estimating future sales for the Pittsburgh Plate Glass Company. A substantial portion of the company's business is the manufacture of windshields and windows for new automobiles. In addition, the company makes glass and paint products used in new construction. Accordingly, you collect the data below:

Year	Net Sales Pittsburgh Plate Glass Company (Millions of Dollars)	Automobile Production (Millions)	Building Contracts Awarded (48 States) (Billions of Dollars)
1948	280.0	3.909	9.43
1949	281.5	5.119	10.36
1950	337.2	6.666	14.50
1951	404.2	5.338	15.75
1952	402.1	4.321	16.78
1953	452.0	6.117	17.44
1954	431.0	5.559	19.77
1955	582.0	7.920	23.76
1956	596.6	5.816	31.61
1957	620.8	6.113	32.17
1958	513.6	4.258	35.09
1959	606.9	5.591	36.42
1960	628.0	6.675	36.58
1961	602.7	5.543	37.14
1962	656.7	6.933	41.30
1963	778.5	7.638	45.62
1964	827.6	7.752	47.38

SOURCE: *Moody's Industrial Manual*, F. W. Dodge Corp.

- a) Find the relationship between net sales and the independent variables automobile production and building contracts awarded, by multiple regression analysis.
 - b) Explain the meaning of the multiple regression equation.
 - c) How well do these variables explain PPG sales?
 - d) Plot the residuals from the multiple regression against the independent variables. Is there any evidence of curvilinearity? Is there any evidence of autocorrelation?
 - e) Forecast 1965 sales on the basis of 9.306 million for auto production and \$49.83 billion for building contracts. If there is evidence of autocorrelation, would this indicate that you should adjust your forecast?
 - f) Compare your estimate from *d* above with the actual Pittsburgh Plate Glass Company sales of \$897.5 million.
8. Refer to the data for Pittsburgh Plate Glass Company sales, automobile production, and building contracts in Problem 7.
- a) Calculate the percent change for each year for the three variables and estimate the multiple regression equation relating percent changes in PPG sales to percent changes in automobile production and building contract awards.
 - b) Explain the meaning of the multiple regression equation.
 - c) Plot the residuals against each of the independent variables. Is there any evidence of curvilinearity? Has the amount of autocorrelation been reduced from that in Problem 7 above?
 - d) Forecast 1965 sales for PPG on the basis of a 20.05 percent increase in auto production and a 5.17 percent increase in building contracts awarded.
 - e) Actual sales of PPG were \$897.5 million. Compare your forecast with this actual value and with the forecast obtained in *d* above.
9. *Note:* This problem requires the use of the matrix multiple regression method (Appendix B to Chapter 23) or else a computer program. Refer to Problem 15 at the end of Chapter 23.
- a) Fit a function of the form $Y_c = a + bX_1 + cX_1^2 + dX_2$ to the data (Y is manufacturing cost; X_1 is production level; X_2 is raw material and labor costs).
 - b) Plot the residuals against the independent variables. Is there any evidence of curvilinearity remaining?
 - c) Is the coefficient c statistically significant?
 - d) Compare the results of this problem with those of Problem 15 in Chapter 23.
10. a) Plot the sales of Sears, Roebuck (Table 19-1, column 2) with disposable personal income (Table 24-2) for the years 1948-1964.
- b) Is there evidence from the data that the relationship may be different for the years 1961-1964 than for the earlier years? If so, graphically estimate the regression lines for the years 1948-1960 and 1961-1964.
 - c) Use the relationship for 1961-1964 to predict sales for 1965, assuming a value of \$465.3 billion for disposable income. Compare the estimate with actual sales of \$6.390 billion.

- d) Find the data on Sears, Roebuck sales and disposable personal income from *Moody's Industrials Manual* and *Survey of Current Business* for the years since 1965. Do the data for these subsequent years confirm a change in relationship between Sears, Roebuck sales and disposable income for the years subsequent to 1961?
11. a) Plot the first differences (percent changes) for Sears, Roebuck sales (Table 19-1, Column 2) and disposable personal income series (Table 24-2) for the years 1948-1964.
- b) Estimate the regression line relating the two series by the graphic method or least squares, as directed. Make an estimate of Sears, Roebuck sales for 1965, assuming a 6.8 percent increase in disposable income.
- c) Note that the years 1962-1964 are all above the regression line, suggesting a possible change in relationship for the latter years. Make a forecast for 1965, assuming that it will have the same deviation from the regression line as 1964 (i.e., run a line through the 1964 point, parallel to the regression line of *b* above, and use this line to forecast 1965 percent change in Sears, Roebuck sales). Compare this forecast with that obtained in *b* above.
- d) In what respect does the method suggested in *c* above differ from part *b* in Problem 10?
12. a) Plot disposable personal income for 1948-1964 on a semilog scale and draw a trend line through the data, as illustrated in Chapter 19. Determine the deviations from the trend line (as a percent of trend) for each year.
- b) Plot the deviations in trend from *a* above with those in Table 19-3 for Sears, Roebuck sales. Calculate the regression line relating the two series.
- c) Forecast Sears sales for 1965, assuming a level of \$465.3 billion for disposable income and a continuation of Sears, Roebuck trend. Compare this with actual sales of \$6.390 billion.
13. a) Compare the methods of forecasting suggested in Problems 10, 11, and 12 above. Which was the most accurate for 1965? Which do you think would be the most accurate in general? Why?
- b) For a long-term projection (say, to 1970), which method would you prefer? Why?
14. a) Estimate the multiple regression between Sears, Roebuck sales, disposable personal income, and time for the period 1948-1964.
- b) Plot the residuals against the independent variables. Is there evidence of curvilinearity? Is there evidence that the relationship may have changed after 1961?
- c) Compare this method of forecasting with those illustrated in Problems 10, 11, and 12. What are the advantages and disadvantages? Is it more useful for long-term or short-term forecasting?
-

15. Some of the variability in Sears, Roebuck sales may be attributable to the fact that many new retail stores are being opened. The number of stores open at the beginning of each year is shown in the table.

NUMBER OF RETAIL STORES—SEARS, ROEBUCK
AT BEGINNING OF FISCAL YEAR

(February 1)

Year	No. of Stores	Year	No. of Stores	Year	No. of Stores
1951	654	1956	709	1961	747
1952	674	1957	721	1962	747
1953	684	1958	732	1963	748
1954	694	1959	736	1964	761
1955	699	1960	741	1965	777
				1966	786

SOURCE: Company annual reports.

- Compute the multiple regression between Sears, Roebuck sales and the independent variables, disposable personal income, and number of stores for the years 1951–1965.
- Plot the residuals against the independent variables. Is there evidence of a change in the relationship after 1961? Explain.
- Is the number of stores statistically significant in explaining Sears, Roebuck sales?

SELECTED READINGS

BRYANT, EDWARD C. *Statistical Analysis*. Rev. ed. New York: McGraw-Hill, 1966.

Chapters 7 and 10 treat simple and multiple regression in a concise manner. Matrix notation is used in the treatment of multiple regression.

CROXTON, FREDERICK E., and COWDEN, DUDLEY J. *Practical Business Statistics*. 3d ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1960, chaps. 24 to 27 and 35 to 37.

Treats a wide variety of topics in simple and multiple correlation.

DRAPER, N. R., and SMITH, H. *Applied Regression Analysis*. New York: John Wiley, 1966.

A detailed treatment of regression analysis at an advanced level. Covers many practical problems, with examples from economics and natural sciences.

EZEKIEL, MORDECAI, and FOX, KARL A. *Methods of Correlation and Regression Analysis*. 3d ed. New York: John Wiley, 1959.

This is the standard book in the field. In the third edition, its major emphasis has been shifted from correlation to regression. Graphic analysis of curvilinear relationships is stressed.

GOLDBERGER, ARTHUR S. *Econometric Theory*. New York: John Wiley, 1964.

An advanced treatment of linear regression, including matrix algebra and systems of simultaneous linear relationships.

HELLWIG, Z. *Linear Regression and Its Application to Economics*. New York: Macmillan, 1963.

A technical treatise, with many examples in economic research.

JOHNSTON, J. *Econometric Methods*. New York: McGraw-Hill, 1963.

A comprehensive study of the linear normal regression model, autocorrelation, and simultaneous equation problems.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics, A New Approach*. New York: The Free Press, 1956, chap. 17.

A thorough, well-organized, easy-to-understand treatment of simple regression.

WILLIAMS, E. J. *Regression Analysis*. New York: John Wiley, 1959.

Provides the practical statistician with a compendium of the classical techniques associated with regression analysis.

25. STATISTICAL QUALITY CONTROL

AMERICAN industry in recent times has adopted a new management technique based on the principles of statistics. This technique is known as statistical quality control or, more simply, quality control. The methods employed had their origin in the 1920's, but World War II led to their widespread adoption by producers of war matériel. Manufacturers were faced with demands for vast quantities of acceptable products in a short time. Specifications were more exacting than ever before. The quality control techniques developed to meet this need proved outstandingly successful in speeding work, reducing manufacturing waste, improving product quality, and bettering product designs. Today, these methods have become an integral and permanent part of management controls.

Quality control methods are applied to two distinct phases of plant operation: (1) the control of a process during manufacture and (2) the inspection of materials to determine their acceptability, whether they be in the raw, semifinished, or completed state. The principal emphasis here will be on the first phase, the control of a process.

TYPES OF VARIATION IN QUALITY

Ordinarily, in a manufacturing process there is a tendency to disregard variation until it causes trouble. If the customer complains of a defective product; if waste, scrap, rejects, or rework increases costs materially; or if sales are lost because a competitor has a more uniform product, a search is instituted in an effort to detect the causes of variability in the product. In the past, and frequently today, such a search has been conducted on the basis of trial and error, and the process is corrected accordingly.

Statistical quality control has demonstrated, however, that such

trial-and-error methods waste time and money (1) because of the lack of a systematic procedure for detection of trouble and (2) because such methods do not become operative until a great many defective parts are discovered by the plant inspector or by the customer. As a consequence, losses are sustained in producing defectives, in excessive inspection costs, in sales, and in goodwill. Manufacturing processes have become so complex and so many things can happen to make them go wrong that it is imperative to have a systematic method of detecting or predicting trouble. Only in this way can prompt corrective action be instituted.

It is evident, then, that any manufacturer must distinguish between permissible variation and excessive variation. His ability to eliminate the latter will be a determining factor in his success or failure.

Statistical quality control permits the partitioning of the total variation of a quality characteristic into two components: (1) *Chance variation* is that which results from many minor causes that behave in a random manner. This type of variation is permissible, and indeed inevitable, in manufacturing. (2) *Assignable variation* is a relatively large variation that can be attributed to special nonrandom causes. It may be excessive in amount so as to require correction. These two types of variation are described below. A *quality characteristic* is simply any measurable variable (such as the thickness of a shingle) or any attribute (such as color) of a part which must be controlled in order that the resulting product be acceptable.

Manufacturing processes are subject to numerous small influences which combine to give a pattern of *chance variation*. This pattern cannot be altered without a change in the process. From time to time other causes of variation enter the process to produce *assignable variation*. Tool wear, a change in the raw material, a new operator, improper machine setting—all can produce assignable variations. The value of quality control lies in its power to detect quickly the assignable variations in a process; in fact, these variations are often discovered before the product becomes defective.

Once the assignable variation in a process has been eliminated by taking corrective action, only the unavoidable chance variation remains. It is possible to measure this chance variation. Then, if the average value of the quality characteristic is set by the engineering specification, it is possible to determine whether the process can conform to these specifications.

CONTROL CHARTS FOR VARIABLES

Control charts are used to distinguish the assignable variation from the chance variation of a process. There are two principal types of

control charts: (1) charts for variables and (2) charts for attributes. As indicated earlier, variables are quality characteristics that can be measured and expressed in numbers, such as the diameter of a bushing. Attributes usually refer to the classification of a quality characteristic into one of two classes, either conforming or not conforming to specifications, as in "accept" or "reject" by visual inspection, or a "go not-go" gauge test. Sometimes quality characteristics which can be measured as variables are actually checked as attributes. Attributes may be judged either by the proportion of units that are defective or by the number of defects per unit. This section is devoted to control of a variable. Control of attributes will be treated later.

Two charts are commonly used in control of variables, the \bar{X} chart and the R (range) chart.

\bar{X} Charts

The \bar{X} chart, or chart for averages, shows variations in the "level" of the process, that is, the arithmetic mean of a quality characteristic being measured. If a process contains no assignable variation in the characteristic controlled, the mean value of the characteristic is the mean of a population of its values, the population being generated by conceiving the process to run *ad infinitum* without change. It is apparent that the actual level of a process cannot be determined, but an accurate estimate of this level can be made by averaging the means of a number of samples say 20 or more (assuming a sample size of 4 or 5). This estimate of the population mean μ is designated \bar{X} .

This same hypothetical population would contain random variation, which may be measured by the standard deviation σ . Since σ , the population standard deviation, is usually unknown, it is necessary to estimate it from data secured by sampling. Such an estimate may be made by the use of either the average range or the average standard deviation of a number of samples. If the sample size is small (about 15 or less), sample range values provide a good estimate of σ . If, however, the sample size is greater than 15, standard deviation values should be employed instead for this purpose.

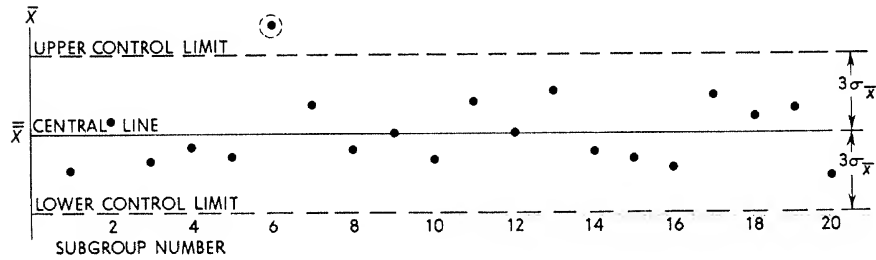
Sample sizes of only 4 or 5 are typical in control charts for \bar{X} . Furthermore, ranges are much easier to calculate than standard deviations. Therefore, R charts are much more commonly used than σ charts in control procedure, and only the former will be included in the following discussion.

The control chart for averages is an excellent application of the distribution of sample means. If, from a population with mean μ and standard deviation σ , all possible random samples of size n are drawn

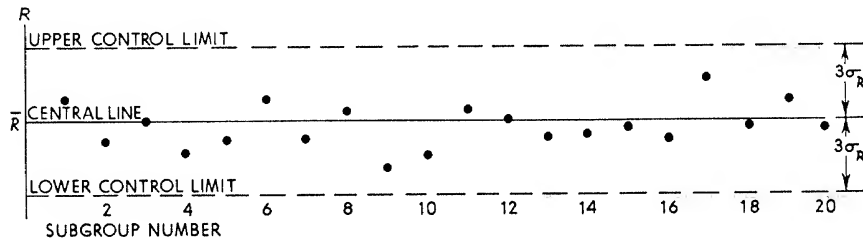
and the averages (arithmetic means) of these samples are placed in a frequency distribution, the resulting distribution will be approximately *normal* with mean μ . The distribution of means is normal even for small samples if the population is normal. Furthermore, the standard deviation of these means (i.e., $\sigma_{\bar{x}}$, the standard error of the mean) will equal σ/\sqrt{n} . The normal distribution pattern permits one to predict the proportion of sample means which will fall within a certain distance of the population mean. In particular, 99.73 percent of the means of

Chart 25-1

CONTROL CHARTS FOR VARIABLES

A. \bar{X} CHART

B. R CHART



samples should fall in an interval defined by $\mu \pm 3\sigma_{\bar{x}}$. This distribution is discussed in greater detail in Chapters 8, 11, and 16.

The distribution of sample means is the foundation of the control chart for averages. When used on a control chart, $\bar{\bar{X}}$, the estimated value of μ , is made the central line and the values $\bar{\bar{X}} + 3\sigma_{\bar{x}}$ and $\bar{\bar{X}} - 3\sigma_{\bar{x}}$ are termed the upper and lower control limits, respectively. The use of 3σ limits is an arbitrary but standard practice for control charts in the United States.

Ordinarily, the value of σ is estimated from a sample, or group of samples, and hence should be represented by the symbol s , as has been done throughout this book. The symbol σ will be used in this chapter, however, in accordance with the almost universal practice among qual-

ity control engineers. Therefore, it should be borne in mind that whenever " σ " is computed from a sample, it is in reality the estimate s , which is subject to sampling errors.

Chart 25-1A is an \bar{X} chart, or control chart for averages. Note that the horizontal scale is designated by subgroup number. In industrial work it is customary to term the sample a "subgroup." Subgroups are samples taken in a certain order. The ordering may be on the basis of time or by lot number or some other plan, but it is important to maintain the order of sampling. The vertical scale is labeled \bar{X} . At the point $\bar{\bar{X}}$ on the vertical scale a horizontal central line is drawn. On either side of this line at a distance of $3\sigma_{\bar{X}}$, parallel dashed lines are drawn. These are the control limits.

R Charts

The R chart shows variations in the ranges of samples. It is similar to the chart for averages in its construction, as shown in Chart 25-1B. The vertical scale is labeled R . A horizontal control line is drawn at \bar{R} , the average of a number of sample ranges. The control limits are dashed and set at a distance of $3\sigma_R$ from the central line, where σ_R is the standard deviation of the sample ranges. (The method of computation will be described later.)

The distribution of the ranges of all possible samples drawn from a normal population is not normal but is skewed in a positive direction. Therefore, as many as 1 percent or more of the cases may exceed the upper $3\sigma_R$ limit. Nevertheless, it works reasonably well to use $3\sigma_R$ limits about the average range \bar{R} as control limits, and this is the usual practice. The chief difficulty is that for small samples the skewness may be so great as to cause the lower control limit ($\bar{R} - 3\sigma_R$) to be negative. In such a case the lower control limit is set at zero, since a range value cannot be negative. If no assignable variation is present in the process, it is expected that practically all the sample range values will fall within the $3\sigma_R$ band about the average range.

Use of Control Charts

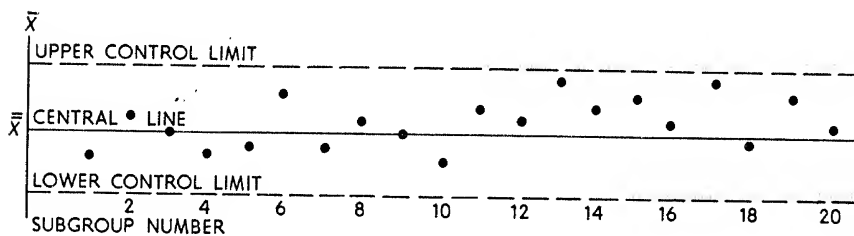
\bar{X} Charts. Assume that the value of the process average μ and its standard deviation σ have been estimated for a certain characteristic—say, the thickness of shingles—and that Chart 25-1A is the control chart for the average value of this characteristic. How shall this chart be used? The general procedure is as follows: Select a sample of the product from the manufacturing process at specified intervals of

time. (The sample size is determined in advance—say, $n = 5$ —and is used in the calculation of the control limits.) Subgroup 1 may have been taken at 8 A.M., subgroup 2 at 8:30 A.M., etc. Calculate the average of each subgroup. Plot these averages on the chart at equal intervals along the horizontal axis. If chance variation *only* is present, virtually all of the sample means should fall inside the control limits defined as $\bar{X} \pm 3\sigma_{\bar{X}}$.

If a point should fall beyond the control limits, the presumption is that assignable causes have affected the process, since the probability of getting such an extreme value by chance is very small. The process is no longer “in control,” but is “out of control.” The importance of ordering the samples is here evident: A point beyond limits indicates that trouble

Chart 25-2

\bar{X} CHART SHOWING SHIFT IN PROCESS AVERAGE



has occurred in the process since the taking of the last sample. The procedure is to investigate immediately to determine the source of this variation. The process may then be shut down until the trouble is located and corrected. Note the average for subgroup 6 on Chart 25-1A, which is out of control, indicating assignable variation.

The use of control charts is an application of the theory of testing hypotheses, described in Chapter 12. The hypothesis is posed that the process average μ is unchanged. When a sample mean falls outside the $3\sigma_{\bar{X}}$ limits, the hypothesis is rejected.

The fact that sample averages follow the normal distribution when assignable variation is absent can be used to detect trouble in a process even though no points may have gone beyond the control limits. With trouble absent, the sample averages should be distributed at random about the central line, with more points near the line than far from it. Then, if an excessively long run—say, 7 points or more—occurs on one side of the central line, as in Chart 25-2, the evidence is that assignable variation has entered the process, causing a shift in process level, even though no points may have fallen beyond the control limits.

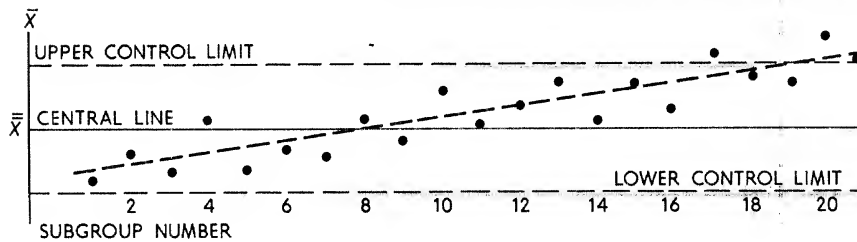
Furthermore, if an upward or downward trend is noted in the points on the average chart, as in Chart 25-3, the evidence also indicates that assignable variation is present. This is frequently the result of uniform tool wear. Thus it is evident that in many cases the control chart for averages, if properly interpreted, can give an indication of impending trouble even though no points have actually exceeded limits. Corrective action can then be taken to avoid production of unsatisfactory items.

R Charts. The plotting of the points on a range chart is similar to that on an \bar{X} chart. The sample range values are plotted at the appropriate subgroup numbers. A point outside limits indicates that the variability of the process has changed and that a search should be instituted immediately to locate the source of the trouble.

The points on a range chart should also be distributed at random in

Chart 25-3

\bar{X} CHART SHOWING INCLINED TREND IN PROCESS AVERAGE



the absence of assignable variation, except that the positive skewness of these distributions means that a few more values should fall below the central line than above it. Any suspicious deviation from such a pattern (even though no points fall outside limits) should be regarded as evidence of a change in the variability of the process.

In summary, control charts for variables provide a basis for action with respect to both the average level and the variability of a process. The charts provide a continuous check on consistency of performance. A proper interpretation of the information on the charts often permits the detection of impending trouble and immediate corrective action.

Why 3-Sigma Limits?

The common use of 3-sigma limits for control charts in this country is rather arbitrary. Theoretically, one should set the limits in each case by balancing the probability of a Type I error (rejecting a true hypothesis, i.e., stopping a process that is running correctly) against the probability of a Type II error (accepting a false hypothesis, i.e., allowing a

faulty process to continue), although the latter probability is difficult to estimate (see Chapter 12). Further, we should find the costs of making the two errors, multiply these costs by the probabilities, and use the expected costs to set control limits (Chapter 9). Finally, we might take the prior probabilities, based on past performance, and revise them by means of current sample evidence and Bayes' Theorem (Chapters 15 and 16) to provide posterior probabilities for use in revising expected costs.

Quality control supervisors might well consider these principles of decision theory in setting control limits for a particular process. For example, if the cost of stopping and checking the process is relatively low, the controls might be tightened by setting the limits less than three standard errors away from the central line.

Control Charts and Specifications

Once a process is brought under control, it is possible to determine whether it is capable of meeting stated specifications. The method is as follows: First, estimate the process dispersion measure σ from the average range \bar{R} (or average standard deviation) of the items sampled. The estimated σ equals \bar{R}/d_2 , where d_2 is a factor found in Table 25-2. Second, on the assumption that the characteristic under control is distributed normally, one can say that nearly all (99.73 percent) of its values should fall within the range $\bar{X} \pm 3\sigma$. This interval of 6σ can then be compared with the tolerance range (upper specification limit minus lower specification limit) to determine whether the process can meet these specifications. Three situations may occur:

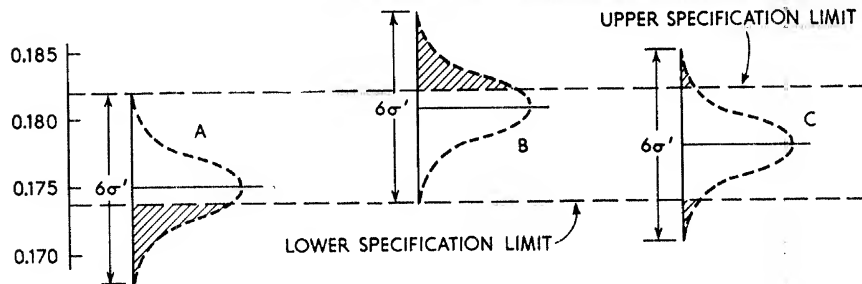
1. If 6σ is greater than the tolerance range, as in Chart 25-4A, the process cannot meet specifications no matter what the level of the process.
2. If 6σ is equal to the tolerance range, as in Chart 25-4B, the process will meet specifications only if the level of the process is midway between the specification limits.
3. If 6σ is less than the tolerance range, as in Chart 25-4C, the process will meet specifications even if the level of the process is allowed to shift within certain limits.

In this way it is possible to judge whether there is excessive variation in the product. Any variation outside specifications is excessive. There are, in general, three corrective actions which may be taken in this case:

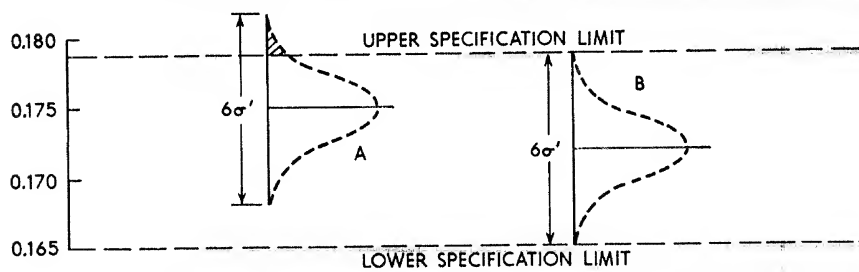
1. Revise the specifications, relaxing tolerances so that the process can meet the new limits.
2. If specifications as written *must* be met, change the process if possible. This may be a minor change, such as resetting a machine or tightening and repairing existing equipment, or it can be an extremely expensive job, involving a change in the raw material, a complete revision of the process, or installation of new machines.

Chart 25-4

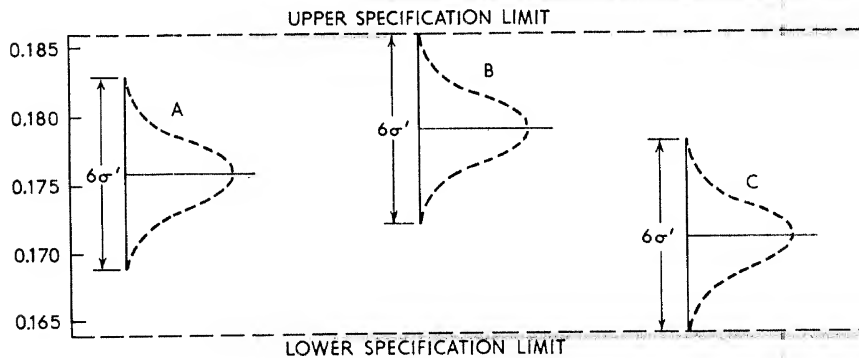
A. PROCESS NOT CAPABLE OF MEETING SPECIFICATIONS



B. PROCESS JUST CAPABLE OF MEETING SPECIFICATIONS



C. PROCESS MORE THAN CAPABLE OF MEETING SPECIFICATIONS



3. If the inspection test is nondestructive, make a 100 percent inspection of the characteristic, and sort the nonconforming from the conforming material. This, too, could be costly and certainly would not assure perfect lots of final product, for it has frequently been demonstrated that 100 percent inspection does not assure perfect segregation.

An Example of Process Control

A capacitor, or condenser used to store an electric charge in a television set, is composed of a ceramic disc which is silvered on each face, attached to two leads, and dipped in a special wax for protection. This example concerns the control of the diameter of the ceramic disc after firing.

Sources of Variation. This process is subjected to numerous sources of variation. Some of these are (1) raw material may vary as a result of its composition, mixing and sizing, drying, or storage; (2) variations may occur in the setting of machines, in level of material in hoppers, in hydraulic pressure applied, and in operation of presses by workers; and (3) kilns may vary in firing time or in temperature.

The most troublesome variations occur in the density of the disc, for wide density variation causes nonuniform shrinkage of the disc when fired. Density is affected by all preliminary operations—particularly the state of the raw material, the level of material in the hopper, and the pressure applied by the press. Also, if the discs are fired too quickly, the rapid rise in temperature causes them to warp, chip, or crack.

Control of the Fixed Diameter of the Disc. For purposes of illustration, assume that this process is just being put under control. Nothing is known about the variability of the process other than that the green discs are controlled by weight before entering the kilns.

The characteristic to be controlled is the fired diameter, which is specified on the drawing as 500 ± 10 thousandths of an inch. The inspector takes 20 subgroups of 5 each and records the readings in thousandths of an inch as deviations from 0.500 inch. (See Table 25-1.)

1. *Calculation of Trial Control Limits.* Add the 20 sample means and divide by 20 to secure the overall mean:

$$\bar{\bar{X}} = \Sigma \bar{X} / n = -2.4 / 20 = -0.12$$

Take this value tentatively as the best approximation of the population mean μ (process level). Now compute the average range from the sample ranges in the same way:

$$\bar{R} = \Sigma R / n = 113 / 20 = 5.65$$

Table 25-1

MEASUREMENT OF FIRED DIAMETER OF CERAMIC DISC

Specification: 500 ± 10 Thousandths of an Inch
 Kiln No. 5—Shift 2 Characteristic: Fired Diameter
 (Deviations from 0.500 Inch in Thousandths of an Inch)

SUBGROUP NUMBER	DISC NUMBER					TOTAL	MEAN	RANGE
	1	2	3	4	5			
1	2	3	4	0	-5	4	0.8	9
2	-3	1	-5	1	1	-5	-1.0	6
3	-2	-1	1	3	-4	-3	-0.6	7
4	-4	-2	1	-3	-4	-12	-2.4	5
5	-1	4	3	-2	6	10	2.0	8
6	3	4	0	1	2	10	2.0	4
7	4	2	4	2	3	15	3.0	2
8	-3	-3	2	2	0	-2	-0.4	5
9	1	2	-3	-2	2	0	0	5
10	1	2	-1	2	-6	-2	-0.4	8
11	-2	2	1	2	1	4	0.8	4
12	-5	-8	-8	0	-4	-25	-5.0	8
13	-2	4	-1	-1	2	2	0.4	6
14	0	-2	-2	-2	1	-5	-1.0	3
15	2	1	1	0	0	4	0.8	2
16	2	0	-4	-5	-1	-8	-1.6	7
17	0	-5	1	-1	-4	-9	-1.8	6
18	-1	0	2	0	1	2	0.4	3
19	2	5	3	-6	2	6	1.2	11
20	-1	-1	3	0	1	2	0.4	4
Total						-12	-2.4	113

The calculation of control limits for the \bar{X} chart requires an estimate of $3\sigma_{\bar{X}}$. Tables have been prepared which simplify this task materially. Enter Table 25-2 at sample size 5 and choose the value of A_2 . It is 0.577. Then $3\sigma_{\bar{X}}$ may be estimated as $A_2\bar{R}$:

$$3\sigma_{\bar{X}} = A_2\bar{R} = 0.577 \times 5.65 = 3.26$$

The upper and lower control limits for the \bar{X} chart are, therefore,

$$UCL_{\bar{X}} = -0.12 + 3.26 = 3.14$$

$$LCL_{\bar{X}} = -0.12 - 3.26 = -3.38$$

The control limits for the *range* chart may be estimated as easily as those for the \bar{X} chart. The upper control limit is $D_4\bar{R}$, where D_4 is found in Table 25-2, for sample size 5. D_4 is 2.114. Then,

$$UCL_R = \bar{R} + 3\sigma_R = D_4\bar{R} = 2.115 \times 5.65 = 11.95$$

Table 25-2

FACTORS USEFUL IN CONSTRUCTION OF CONTROL CHARTS*

NUMBER OF ITEMS IN SAMPLE	CHART FOR AVERAGES, Factors for Control Limits	CHART FOR RANGES		
		Factors for Central Line	Factors for Control Limits	
n	A_2	d_2	D_3	D_4
2	1.880	1.128	0	3.267
3	1.023	1.693	0	2.575
4	.729	2.059	0	2.282
5	.577	2.326	0	2.115
6	.483	2.534	0	2.004
7	.419	2.704	0.076	1.924
8	.373	2.847	0.136	1.864
9	.337	2.970	0.184	1.816
10	.308	3.078	0.223	1.777
11	.285	3.173	0.256	1.744
12	.266	3.258	0.284	1.716
13	.249	3.336	0.308	1.692
14	.235	3.407	0.329	1.671
15	.223	3.472	0.348	1.652

* Note: These factors assume a normal distribution, with true value of σ known.

SOURCE: American Society for Testing Materials, *Manual on Quality Control of Materials*, Table B2, p. 115. For more detailed table and explanation, see Acheson J. Duncan, *Quality Control and Industrial Statistics* (3d. ed.; Homewood, Illinois, Richard D. Irwin, 1965), Table M, p. 927.

Similarly, the lower control limit is $D_3\bar{R}$, where D_3 is 0 in Table 25-2:

$$LCL_R = \bar{R} - 3\sigma_R = D_3\bar{R} = 0 \times 5.65 = 0$$

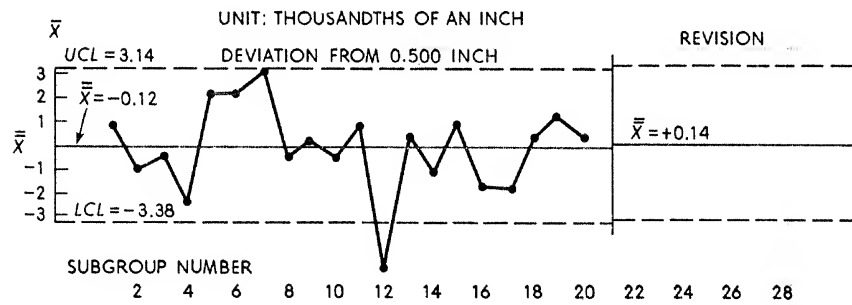
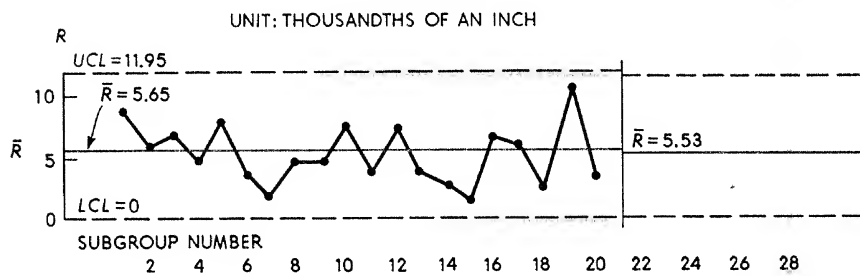
Here, because of the small sample size, the computed value of LCL_R is placed at zero.

2. *Interpretation of the Charts.* In the chart for averages (Chart 25-5A) all points are within the control limits except subgroup 12. No trend is apparent, and there is no indication of excessively long runs. It is concluded that, with the exception of subgroup 12, the process is free of assignable variation. (In this case, an investigation disclosed that the sagger from which subgroup 12 was drawn had been red-tagged, that is, rejected, because it did not meet green-density standards but had been processed through error.) The chart for ranges (Chart 25-5B) also shows an "in control" condition with respect to process variability.

3. *Revision of Limits.* Since the \bar{X} chart contains a subgroup outside limits, it would not be proper to use the value of $\bar{\bar{X}} = -0.12$ as the

Chart 25-5

CONTROL CHARTS FOR FIRED DIAMETER OF CERAMIC DISCS

A. \bar{X} CHARTB. R CHART

best estimate of the process level (average) under control. As a better approximation, eliminate subgroup 12 and compute a revised $\bar{\bar{X}}$ from the remaining 19 groups:

$$\bar{\bar{X}}_{\text{rev}} = \frac{\sum \bar{X}}{n} = \frac{-2.4 - (-5.0)}{20 - 1} = \frac{+2.6}{19} = +0.14$$

Although the range chart shows control, it will give a better estimate of normal process variability if subgroup 12 (from the rejected sagger) is eliminated. Revised values of \bar{R} and $A_2\bar{R}$ for the remaining 19 groups are

$$\bar{R}_{\text{rev}} = \frac{\sum R}{n} = \frac{113 - 8}{20 - 1} = \frac{105}{19} = 5.53$$

$$A_2\bar{R}_{\text{rev}} = 0.577 \times 5.53 = 3.19$$

Revised control limits for the \bar{X} chart are

$$UCL_{\bar{X}} = 0.14 + 3.19 = 3.33$$

$$LCL_{\bar{X}} = 0.14 - 3.19 = -3.05$$

Revised control limits for the R chart are

$$UCL_R = \bar{R} + 3\sigma_R = D_4\bar{R} = 2.115 \times 5.53 = 11.70$$

$$LCL_R = \bar{R} - 3\sigma_R = D_3\bar{R} = 0 \times 5.53 = 0$$

The revised central lines and control limits are drawn on the right side of Chart 25-5. The points on the two charts still lie within the new control limits, except for subgroup 12, which is expected to fall outside the limits on the \bar{X} chart. The revised values of \bar{X} and \bar{R} are the best estimates of the true process average and range which are possible on the basis of 20 subgroup readings. As additional data are secured, it may be desirable to revise these estimates.

4. *Ability of Process to Meet Specifications.* Since the R chart exhibits control, it is possible to estimate the value of σ , the process variation measure, as follows:

$$\sigma = \frac{\bar{R}}{d_2} = \frac{5.53}{2.326} = 2.38$$

where d_2 is a factor secured for subgroup size 5 in Table 25-2. The range 6σ is then $6 \times 2.38 = 14.28$. According to specifications, the tolerance range is 20. Since 6σ is less than the tolerance range, this process can meet the specifications if the process level is satisfactory. If the characteristic follows the normal pattern, nearly all of the fired diameters will fall between $\bar{X} \pm 3\sigma$ or 0.14 ± 7.14 . Specifications are 0 ± 10 . It is evident that this process will meet specifications if it is controlled at the present level.

5. *Future Use of Charts.* For the next period, the two control charts will have new central lines and new control limits, as indicated above. The inspector will compute and plot the values of \bar{X} and R immediately upon measuring the five members of the subgroup. In this way, he can detect trouble promptly and undertake an investigation at once to determine the cause.

In summary, the ceramic disc diameters are shown to be adhering to a level of $+0.14$ thousandths (specification: 0) with a variation of 2.38 (σ), on the basis of the first 20 subgroups. This process seems capable of control and will meet specifications if controlled at the present level.

CONTROL OF ATTRIBUTES

As mentioned earlier, the control of *variables* employs the \bar{X} and R chart technique. Control of *attributes* is achieved by use of either the p

chart or the c chart, the first for proportion of units that are defective and the second for number of defects per unit. A defect is any imperfection which will render the article or part unfit for the purpose originally intended. For example, if white enamel panels are inspected visually, a chip, black spot, crack, imperfect coverage of enamel, off color, or bend in the panel is an imperfection which will cause the article to be rejected and is consequently termed a defect. If a panel has one or more of these defects, it is counted as one *defective* panel, while a count must be made to determine the number of *defects*.

Fraction Defective Chart or p Chart

The p chart is used to control the *proportion of units that are defective* in a given attribute. This chart has its theoretical basis in the binomial distribution and generally gives best results when the sample size is large—say, at least 50.

The central line is placed at \bar{p} , the average fraction defective, where \bar{p} is the number of defectives divided by the total number inspected. The control limits are $3\sigma_p$ from the central line, where $\sigma_p = \sqrt{\bar{p}\bar{q}/n}$ for sample size n , and $\bar{q} = 1 - \bar{p}$. As in the case of the \bar{X} chart, the value of \bar{p} is subject to revision as more data are secured. The following case illustrates the application of this chart.

An inspection procedure in the manufacture of spark plugs calls for an inspection for defectives on finished plugs in lots of 200 each. The check is visual and can be made rapidly by experienced operators. The data in Table 25-3 show the number of defectives found in the inspection of 24 lots of 200 each. The computations are as follows:

Average fraction defective:

$$\bar{p} = \frac{192}{4,800} = 0.040$$

$$\sigma_p = \sqrt{\frac{\bar{p}\bar{q}}{n}} = \sqrt{\frac{0.040 \times 0.960}{200}} = 0.0138$$

$$3\sigma_p = 3 \times 0.0138 = 0.041$$

$$\text{Upper control limit: } \bar{p} + 3\sigma_p = 0.040 + 0.041 = 0.081$$

$$\text{Lower control limit: } \bar{p} - 3\sigma_p = 0.040 - 0.041 = -0.001$$

LCL set at 0.

Chart 25-6 is the control chart for these lots. Note that there is one point above the upper control limit, indicating one lot which had more

Table 25-3

INSPECTION DATA ON COMPLETED
SPARK PLUGS(4,800 SPARK PLUGS IN 24 LOTS
OF 200 SPARK PLUGS EACH)

Lot Number	Number Defectives	Fraction Defective
1	10	0.050
2	7	0.035
3	14	0.070
4	4	0.020
5	20	0.100
6	11	0.055
7	14	0.070
8	8	0.040
9	6	0.030
10	12	0.060
11	15	0.075
12	5	0.025
13	8	0.040
14	6	0.030
15	10	0.050
16	13	0.065
17	7	0.035
18	5	0.025
19	3	0.015
20	4	0.020
21	1	0.005
22	3	0.015
23	2	0.010
24	4	0.020
Total	192

defectives than expected. The last eight lots are all below the central line, indicating that the fraction defective has changed to a lower level during this period. If this trend continues, it will be desirable to revise the value of \bar{p} and to establish new, closer control limits. The introduction of a p chart frequently results in a rapid decrease in number of defectives, since it sounds the alarm for immediate action in case of trouble.

In most cases, the number of items inspected varies from lot to lot, causing the upper and lower control limits to vary. Although this requires more computations, the interpretation of such a chart is precisely the same as one with constant control limits.

Chart 25-6

p CHART FOR SPARK PLUG INSPECTION
(24 LOTS OF 200 SPARK PLUGS EACH)

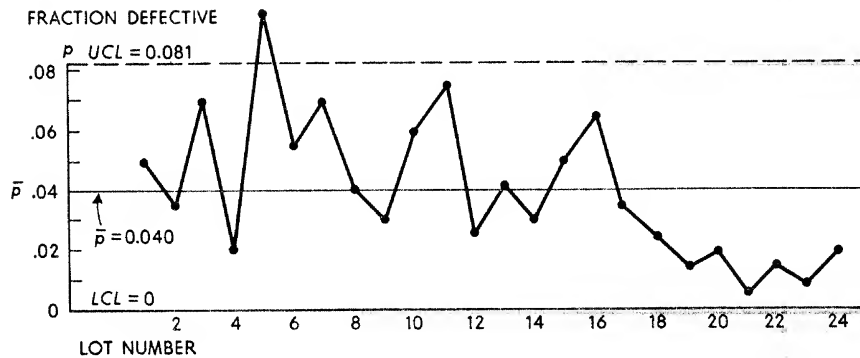


Chart for Number of Defects per Unit or *c* Chart

This chart is employed to control the actual *number of defects per unit*, rather than the number of defective units. The theoretical basis of the chart is the Poisson distribution. The *c* chart is most frequently used where (1) a natural unit does not exist, as in defects per 100 square yards of cloth, the unit of area being arbitrary, or (2) where the unit is quite complex (e.g., aircraft instruments), so that almost all units have some defects. The so-called area of opportunity (e.g., 100 square yards of an identical type of cloth) for the occurrence of a defect must be held constant from part to part for this chart to be effective as a control. The number of units inspected, however, may still vary from sample to sample.

The *c* chart is similar to the *p* chart in its construction and interpretation. The central line is placed at \bar{c} , the average number of defects per unit. The upper and lower control limits are placed at $\bar{c} \pm 3\sigma_c$, where $\sigma_c = \sqrt{\bar{c}}$. Special tables are not needed to calculate these limits.

ACCEPTANCE SAMPLING

The principles of quality control have been applied above to the regulation of the manufacturing process itself. Another important field of quality control is acceptance sampling. As its name implies, acceptance sampling is a procedure for sampling a lot in order to determine whether to accept it as conforming to standards or to reject it. If rejected, it may be submitted to 100 percent inspection or returned to the supplier. A purchaser may wish to sample the quality of a shipment of

goods received or a manufacturer may submit his own output to acceptance sampling at various stages of production. The purpose of acceptance sampling, therefore, is to determine whether to accept or reject a product. It does not attempt to control quality during the manufacturing process, as do the techniques described earlier in the chapter.

It is often preferable to inspect only a sample, rather than the entire lot, to determine its acceptability. This is particularly true when inspection is very costly or destructive. Even if 100 percent inspection is feasible, a carefully worked out sampling plan may produce equally good or better results at lower cost. An acceptance sampling plan will improve the quality of the product through rejecting defective lots and bringing pressure to bear on suppliers to improve quality. The lot can be judged promptly, with a known probability of making a mistake.

While the theory is complex, acceptance sampling is simple in practice and can be applied by inspectors without advanced statistical training. The techniques will not be described here but may be found in Eugene L. Grant, *Statistical Quality Control* (3d ed.; New York: McGraw-Hill, 1964), Part III; Acheson J. Duncan, *Quality Control and Industrial Statistics* (3d ed.; Homewood, Illinois: Richard D. Irwin, 1965), Parts II and III; or Dudley J. Cowden, *Statistical Methods in Quality Control* (Englewood Cliffs, New Jersey: Prentice-Hall, 1957), Chapters 30 to 40.

The three principal types of acceptance sampling plans now in use are as follows:

1. The single-sampling plan specifies the sample size and the number of defective units in the sample that will cause the entire lot to be rejected. If a smaller number of defectives is found, the lot is accepted.
 2. In a double-sampling plan, a smaller sample can be taken to begin with. If it contains a specified number c_1 or fewer defective units, the lot is immediately accepted; if it contains more than c_2 , a larger number, the lot is rejected. In the intermediate case, however, a second larger sample is taken. Then, if the combined number of defectives in the two samples is c_2 or less, the lot is accepted; otherwise, it is rejected. Double sampling is preferable to single sampling in reducing the total amount of inspection on very good or very poor lots that can be judged on the first sample. It also has the psychological advantage of giving a tentatively rejected lot a second chance. When many second samples are required, however, double sampling may be more complicated and expensive than single sampling.
 3. In sequential sampling, the size of sample is not determined in
-

advance. Instead, a decision is made after each observation or group of observations to (1) accept, (2) reject, or (3) suspend judgment and continue sampling until a decision is ultimately reached. Sequential methods permit reaching a decision on the basis of even fewer observations than other plans in the case of very good or very bad lots, but the procedure is relatively complex in operation.

SUMMARY

Quality control is an application of hypothesis testing which has come into widespread use in recent years, both for the control of a process during manufacture and for determining the acceptability of a product. This chapter is primarily devoted to process control.

All products vary in quality. Control charts are used to separate the normal *chance* variation from *assignable* variation (attributable to non-random causes) so that the latter can be promptly recognized and remedied. The principal types of control charts are for *variables*, or measurable characteristics, and for *attributes*, or traits that are either present or absent (e.g., passing a "go not-go" gauge test) or nonmeasurable (e.g., color).

The \bar{X} chart for variables is used to control the average value or "level" of a quality characteristic. To construct an \bar{X} chart, draw horizontal lines at the estimated population mean on the vertical scale and at $3\sigma_{\bar{X}}$ control limits on either side. These limits are usually estimated from the average of sample ranges. Plot subgroup averages at equal intervals along the horizontal axis.

The R chart for variables is used to control the variability of the process. To construct an R chart, draw horizontal lines at \bar{R} and at the $3\sigma_R$ limits. If the lower control limit is negative, place it at zero. Then plot the subgroup ranges as in the \bar{X} chart.

Nearly all of the points should fall within the control limits of an \bar{X} or R chart if chance variation alone is present. If a point falls outside the limits or if about seven or more consecutive points fall on one side of the central line or if they show an upward or downward trend, assignable variation is probably present. This should be corrected promptly.

Control limits may be set at intervals other than three standard errors from the central line by application of the decision-theory principles described in Chapters 9, 12, 15, and 16.

The range, estimated as 6σ , must be less than or equal to the specified tolerance range, as shown in Chart 25-4, for the process to meet specifications. If it cannot, the manufacturer can revise specifications, change

the process, or resort to 100 percent inspection unless inspection is too costly or destructive.

The example of the ceramic disc illustrates how process control works—that is, how to calculate trial control limits with the aid of tables, how to interpret the charts, revise the limits, and gauge the ability of the process to meet specifications.

The control of attributes is achieved through the use of either p charts for proportion of units that are defective or c charts for number of defects per unit. The latter are used where no natural unit exists or where the unit is so complex that virtually all units have some defects. These charts are constructed and interpreted in much the same way as the control charts for variables summarized above.

Acceptance sampling is an economical and efficient method of determining whether to accept or reject a shipment or stock of material, based on a sample. This may be a single or double sample or a sequential plan in which the amount of sampling depends on the results of successive tests. Quality control and acceptance sampling have come into widespread use in industrial management, since they help produce a better product at a lower cost.

PROBLEMS

1. Distinguish between:
 - a) Process control and acceptance sampling.
 - b) Chance variation and assignable variation.
 - c) Variables and attributes, as applied to quality characteristics.
 - d) \bar{X} and R charts for variables.
 - e) p and c charts for attributes.
2.
 - a) Describe two situations in which the pattern of points on a control chart would indicate trouble even if no points actually fall outside control limits.
 - b) Explain how to determine whether or not a process is capable of meeting specifications.
 - c) If a process cannot meet specifications, what corrective action can be taken?
3. One of the critical component parts of a product manufactured by your company is a size $\frac{5}{16}$ in. carbon steel bolt. In order to meet product specifications this bolt must have a hardness rating between 77.5 and 89.5 on the Rockwell "B" Hardness Scale. Following a heat treatment designed to produce the desired hardness, a sample of four bolts is drawn at random from each lot, and each bolt is tested for hardness. Ten of these samples, taken in consecutive order, test on the Rockwell "B" Scale as shown in the table.¹

¹ Ten samples are used here to minimize computations. In practice, however, at least 20 or 25 samples are needed for reliable results.

1	2	3	4	5
85.0	87.0	82.0	82.5	89.0
84.5	81.0	93.0	83.0	81.5
85.0	80.5	85.0	85.0	82.0
87.0	79.0	84.5	82.5	84.0
6	7	8	9	10
83.0	84.5	89.0	85.5	89.0
89.0	85.0	88.0	89.5	85.5
83.0	85.0	85.0	89.0	87.0
81.5	88.0	83.5	82.5	89.0

- Set up \bar{X} and R charts to control the hardness of these bolts. Show all calculations, and plot results.
 - Does the heat-treating process appear to be in statistical control? If so, what is your best estimate of the average hardness rating of all bolts produced by this process?
 - If any points are out of control, revise the limits accordingly, and plot the results on the charts.
 - Can this process meet specifications? Explain.
4. Following are mean net weights (expressed as deviations from 1,000 grains) and ranges, both in grains, of 20 subgroups, each consisting of five bottles of sodium bicarbonate. These are filled by machine and labeled "100 ten-grain tablets."

Subgroup Number	\bar{X}	R	Subgroup Number	\bar{X}	R
1	4.6	5	11	0.4	2
2	4.4	3	12	8.0	6
3	4.0	9	13	2.2	5
4	5.0	6	14	5.6	13
5	0.8	2	15	7.2	11
6	2.4	9	16	2.2	8
7	7.2	10	17	4.6	5
8	4.4	4	18	-1.8	6
9	1.8	8	19	7.4	6
10	3.2	11	20	6.0	4

- Set up \bar{X} and R charts to control the operation of the bottle-filling machine. Show all calculations and plot results.
 - Is this process in control? Cite evidence to support your conclusion.
 - If any points are out of control, revise the limits accordingly, and plot the results on the charts.
 - If specifications are 4 ± 8 (i.e., tolerance range 16), can this process meet specifications? Explain.
5. Following are the number of defective electric-shaver motors inspected during each of 23 working days of October, in daily samples of 100.

October	Number Defective	October	Number Defective	October	Number Defective
1	5	11	2	23	2
2	9	12	1	24	3
3	10	15	2	25	5
4	10	16	2	26	5
5	13	17	3	29	3
8	10	18	6	30	2
9	13	19	3	31	3
10	2	22	3		

- a) Construct a p chart to control the quality of the motors.
- b) It is reported that a faulty machine used in assembling the motors was repaired during the month. If there is evidence that the fraction defective changed to a lower level during this period, discard earlier observations, revise \bar{p} , compute closer control limits, and plot the results for future use.
6. A test of 2,000 transistors, in 20 lots, each containing 100 transistors, shows 10 percent defective on the average. What is the maximum percent defective the inspector should allow on the next lot for it to be within $3\sigma_p$ control limits?
7. A quality control engineer is about to set up a control chart for a production process. The process, when in control, produces items with a mean of 40 and a standard deviation of 5. For simplicity, we assume that there are two states in which the process is out of control, one with a process mean of 48 and the other with a process mean of 36. Both have a process standard deviation of 5 (there is never any change in the variability of the process). The costs (economic losses) for these various events are shown in the table.

Possible Events:		
Process Average Is	Action: Accept the Process	Action: Reject the Process
36	\$ 800	\$ 0
40	0	1,200
48	1,000	0

The quality control engineer wants to use an \bar{X} Chart, sample size 4, having control limits $40 \pm k\sigma_{\bar{X}}$. He wishes to select an optimal value for k . Accordingly, he constructs the following table:

Process Average Is	Average (Expected) Costs		
	$k = 1$	$k = 2$	$k = 3$
36	A	B	C
40	D	E	F
48	G	H	I

- a) Find the values A through I to fill in the table.
- b) Explain how you might go about deciding what value of k to use.

SELECTED READINGS

BOWKER, ALBERT H., and LIEBERMAN, GERALD J. *Engineering Statistics*. Englewood Cliffs, New Jersey: Prentice-Hall, 1959.

Contains a readable and authoritative treatment of quality control.

COWDEN, DUDLEY J. *Statistical Methods in Quality Control*. Englewood Cliffs, New Jersey: Prentice-Hall, 1957.

A broad treatise on statistical techniques in process control and product control, including economic considerations, with a minimum of mathematics in the text sections.

DUNCAN, ACHESON J. *Quality Control and Industrial Statistics*. 3d ed. Homewood, Illinois: Richard D. Irwin, 1965.

A broad survey of statistics useful in industrial research. Parts II and III cover acceptance sampling, and Part IV covers control charts.

GRANT, EUGENE L. *Statistical Quality Control*. 3d ed. New York: McGraw-Hill, 1964.

A very readable, nontechnical working manual, designed for use by production and inspection supervisors, engineers, and management.

JURAN, J. M. (ed.). *Quality Control Handbook*. 2d ed. New York: McGraw-Hill, 1962.

An encyclopedia on managerial aspects of quality control and applications of quality control principles in specific industries, with less emphasis on statistical techniques.

•

APPENDIXES

A. GLOSSARY OF SYMBOLS

\sim	Not.
\approx	Approximately equals.
$— $	Given
$ $	Absolute value of enclosed symbol, ignoring sign.
$!$	Factorial: $n! = 1 \times 2 \times 3 \times \cdots \times n$.
$>$	Greater than.
\geq	Greater than or equal to.
$<$	Less than.
\leq	Less than or equal to.
a, A	Value of Y_c in trend or regression equation when all X 's = 0; a is for sample, A is for population.
A'	Transpose of matrix A .
A^{-1}	Inverse of square matrix A .
α	Average number of customers served per unit of time, i.e., service rate (alpha).
b, B	Slope of trend line; slope of higher degree curve at Y axis; simple regression coefficient, where b is for sample, B is for population.
b_1, B_1 , etc.	Net regression coefficients in multiple regression; where b_1, b_2, \dots is for sample, B_1, B_2, \dots is for population.
β_1, β_2, \dots	Standardized values (betas) of b_1, b_2, \dots .
C	Number of combinations; cyclical component in time series, expressed as percent of trend (T).
$C(n)$	Cost of sample of size n .
c	Cost per unit; constant determining curvature in second-degree equation; defects per unit, in quality control.
D	Factor used in determining unit normal loss function (Appendix E).
d	Deviation of class midpoint from assumed mean of frequency distribution in class interval units.
df	Degrees of freedom.
E	Expected value.
EMV	Expected monetary value.

ENGS	Expected net gain from sampling.
EOL	Expected opportunity loss.
EVPI	Expected value of perfect information.
EVSI	Expected value of sample information.
e	The constant 2.718
F	Cumulative frequency for all classes below median or quartile interval.
f	Frequency or number of items in any class ($\Sigma f = n$)
f.p.c.	Finite population correction.
G	Geometric mean.
I	Irregular component in time series expressed as percent of $T \times C \times S$; identity matrix ($I = A \times A^{-1}$).
i	Class interval in a frequency distribution; subscript denoting the i th item in a set of items.
K	Break-even value.
k	Number of constants in an equation; sampling interval in systematic selection; number of replicate samples drawn from population.
L	Lower limit of class containing median or quartile in a frequency distribution.
LCL	Lower control limit in quality control chart.
$L_N(D)$	Unit normal loss function (find D in Appendix E).
$L(X)$	Opportunity loss.
l_o, l_u	Opportunity loss per unit from overstocking (l_o) or understocking (l_u).
log	Logarithm.
M	Number of secondary units in population—cluster sampling.
M_i	Number of items in i th stratum of stratified sample; number of secondary units in i th primary unit in cluster sampling.
MCD	Months for cyclical dominance: $MCD = \bar{I} / \bar{C} $.
M.D.	Mean (average) deviation.
Md	Median.
M_0, M_1	Mean of decision maker's prior (M_0) or posterior (M_1) betting distribution about unknown mean μ .
m	Mean (or variance) of Poisson distribution; number of independent variables in matrix solution of multiple regression.
m_i	Sample size of i th stratum in stratified sample; number of secondary units sampled in i th primary unit, in cluster sample.
μ	Arithmetic mean of a population (μ).
μ_h	Hypothetical population mean.
N	Number of items in a population; number of primary units in population—cluster sampling.
n	Number of items in a sample; number of primary units in cluster sample.
n'	Number of individuals in queue, excluding those being served.
n	Subscript for a given year, index numbers (base period has subscript zero).
OC	Operating characteristic.
P	Probability.

$P(A)$	Probability that event A will occur.
$P(A B)$	Conditional probability of A given B.
$P(A, B)$	Joint probability of A and B.
P_0, P_n	Probability of no one (P_0) or of n individuals (P_n) in queue.
$P(p)$	Prior probability that fraction defective is p .
p	Population proportion; probability of a success; price.
p_h, q_h	Hypothetical population proportions.
p_s	Sample proportion.
π	The constant 3.14159 . . . ; profit (π).
Q	Quartile deviation.
Q_1, Q_3	First and third quartiles ($Q_2 = \text{median}$).
q	Population proportion, probability of a failure, where $q = 1 - p$; quantity (e.g., units stocked).
q_s	Sample proportion, where $q_s = 1 - p_s$.
q^*	Optimal quantity of units stocked.
R	Ratio; range; coefficient of multiple correlation.†
\bar{R}	Arithmetic mean of several sample ranges.
R^2	Coefficient of multiple determination.†
r	Coefficient of simple correlation;† number of successes (e.g., defectives) in sample.
r^2	Coefficient of simple determination.†
r_s	Coefficient of simple correlation for a sample.
S	Seasonal index.
S_k	Coefficient of skewness.
S_0, S_1	Standard deviation of decision maker's prior (S_0) or posterior (S_1) betting distribution about unknown mean μ .
S_{YX}	Standard error of estimate in simple regression.
S^2_{YX}	Unexplained variance in simple regression.
$S_{Y \cdot 12} \dots$	Standard error of estimate in multiple regression.
$S^2_{Y \cdot 12} \dots$	Unexplained variance in multiple regression.
S^2_*	Reduction of prior variance as a result of taking sample.
Σ	Sum, total (capital sigma).
s	Standard deviation.†
s^2	Variance.†
$s_{\bar{X}}$ etc.	Standard error of the mean;† s is used with other subscripts for standard errors of other measures.
S_{Y-Y_c}	Standard error of an individual forecast.
$s^2_{Y_c-\bar{Y}}$	Explained variance.
σ	Standard deviation (small sigma) of a population; or, in quality control, its estimated value.
$\sigma_{\bar{X}}$, etc.	Standard error of the mean; σ is used with other subscripts for standard errors of other measures.
T	Total (population); trend ordinate in time series ($T = Y_c$).
T_i	Total estimated for i th cluster, in cluster sampling.
t	Deviation of sample mean, etc., from population mean, expressed in standard error units. The t distribution applies to small samples. Slope of opportunity loss function.

† Population value as estimated from a sample (except computer printout of R in Chapter 23, which is not adjusted for sample bias).

UCL	Upper control limit in quality control chart.
u	Standard normal deviate: $u = (X - \mu)/\sigma$ in normal distribution; utility value.
w, w'	Waiting time in queue, including (w) or excluding (w') time being serviced.
w_i	Weight of i th stratum in stratified sampling.
X, Y	Independent and dependent variables measured from zero.
\bar{X}, \bar{Y}	Arithmetic means of X and Y in a sample. (Subscripts 1, 2, etc., refer to different samples.)
$\bar{\bar{X}}$	Arithmetic means of several sample means.
\bar{X}_a	Assumed mean of X .
X_1, X_2, \dots	Independent variables in multiple regression.
x, y , etc.	Variables measured from means; e.g., $x = X - \bar{X}$, $y = Y - \bar{Y}$.
Y	Dependent variable in trend or regression analysis.
Y'	Value of Y adjusted for X_1 in graphic method of multiple regression.
Y_c	Value of Y computed from trend or regression equation.
\bar{Y}_c	Population mean estimated from cluster sample.
\bar{Y}_i	Sample mean of i th stratum in stratified sample; sample mean of secondary units in i th primary unit of cluster sample.
\bar{Y}_j	Mean of replicate sample.
\bar{Y}_R	Ratio estimate of true mean μ_Y .
\bar{Y}_s	Estimate of overall mean from stratified sample.
z	Deviation of actual value of Y from computed value Y_c : $z = Y - Y_c$.

B. LOGARITHMS

HOW TO USE THE TABLE OF LOGARITHMS

LOGARITHMS are used to simplify the operations of multiplication, division, raising numbers to powers, and extracting roots. They are especially valuable in constructing ratio charts, in computing the geometric mean, and in fitting certain types of secular trend curves.

The common logarithm of a number is the power of 10 which is equal to that number. For example, the third power of 10 is 1,000, so

$$\log 1,000 = 3$$

That is, the logarithm of 1,000 is 3 because $10^3 = 1,000$. Similarly, $\log 100 = 2$, $\log 10 = 1$, $\log 1 = 0$, $\log 0.1 = -1$, $\log 0.01 = -2$, etc. For intermediate numbers the logarithm is a whole number, as above, followed by a decimal fraction.

The whole number part of a logarithm (to left of the decimal point) is called the *characteristic*, and the fractional part (to the right of the decimal point) is called the *mantissa*. To find the logarithm of any number, determine the characteristic from the following rules and look up the mantissa in the accompanying table.

Rules for Determining the Characteristic

1. The characteristics of the logarithms of all numbers greater than one are positive, and their numerical values are one unit less than the number of digits to the left of the decimal point in the numbers themselves.

Examples:

<i>Number</i>	<i>Characteristic of Logarithm</i>
286.	2
12,769.	4
1,008.73	3
1.827.	0

2. The characteristics of the logarithms of all numbers between zero and one are negative, and their numerical values are one unit greater than the number of zeros between the decimal point and the first significant digit of the numbers themselves. A negative value is indicated either by a minus sign written above the characteristic or as a positive number followed by -10 , as shown below.

Examples:

<i>Number</i>	<i>Characteristic of Logarithm</i>
0.764	$\bar{1}$, or 9... -10
0.031	$\bar{2}$, or 8... -10
0.02793	$\bar{2}$, or 8... -10
0.00004	$\bar{5}$, or 5... -10

3. The number zero and negative numbers have no logarithms.

How to Find the Mantissa

The following table (pp. 692-93) shows four-place mantissas of logarithms for three-digit numbers. This table is accurate enough for most business and economic data. For convenience in printing, decimal points are omitted, but each entry in the table must be interpreted as a four-place decimal. Mantissas are always positive. The mantissa of any number of three digits or less can be read directly from the table. The first two digits are found in the column labeled "N" at the left of the page, and the third digit is found at the top of the page. Thus, to find the logarithm of 316, write down the characteristic 2 from Rule 1 above, followed by the mantissa .4997 from the table. This is found by moving down the column on the left to 31 and going to the right under the column headed 6. The log of 316 therefore is 2.4997.

Examples:

$$\begin{aligned}\log 3.160 &= 0.4997 \\ \log 0.316 &= \bar{1}.4997, \text{ or } 9.4997 - 10 \\ \log 180,000 &= 5.2553 \\ \log 0.031 &= \bar{2}.4914, \text{ or } 8.4914 - 10\end{aligned}$$

The logarithms of four-place numbers may be determined by interpolation. Thus, to find the log of 3.162, go two tenths of the way from log 3.160 (i.e., 0.4997) to log 3.170 (i.e., 0.5011). This is $0.4997 + 0.2 \times 0.0014 = 0.5000$.

Antilogarithms

To find the antilogarithm or natural number corresponding to a logarithm, find the nearest logarithm in the table and read the first two digits of the corresponding natural number from the left-hand column and the third digit from the top row. Thus, to get the antilog of 3.3101, find 3096, the nearest mantissa in the table, and read across and up to the number 204. Then from the rules on characteristics the answer is 2,040. This value may also be interpolated if four-place accuracy is desired.

Rules for Using Logarithms

1. To multiply numbers add their logarithms. Then look up the antilogarithm of their sum. The fact that numbers may be multiplied by adding their logarithms is the most basic property of logarithms.

Example: Multiply 19 by 28:

$$\begin{array}{rcl} \log 19 & = & 1.2788 \\ \log 28 & = & 1.4472 \\ \hline \log \text{ product} & = & 2.7260 \\ \text{product} & = & \text{antilog } 2.7260 = 532 \end{array}$$

2. To divide one number by another, subtract the logarithm of the latter from that of the former. Then look up the antilogarithm of the difference.

Example: Divide 532 by 28:

$$\begin{array}{rcl} \log 532 & = & 2.7259 \\ \log 28 & = & 1.4472 \\ \hline \log \text{ difference} & = & 1.2787 \\ \text{quotient} & = & \text{antilog } 1.2787 = 19.0 \end{array}$$

3. To raise a number to a given power, multiply the logarithm of the number by the exponent of the power and look up the antilogarithm of the product.

4. To extract any root of a number, divide its logarithm by the index of the root and look up the antilogarithm of the quotient.

FOUR-PLACE LOGARITHMS

N	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6336	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396

FOUR-PLACE LOGARITHMS (*Continued*)

N	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996

C. SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000

HOW TO FIND A SQUARE ROOT

SQUARE ROOTS can be read from the following table by any of three methods:

1. For any whole number from 1 to 1,000, listed in the N column, find the square root in the same row of the \sqrt{N} column. Thus, the square root of 458 (in the N column) is 21.4+ (in the \sqrt{N} column).
2. For any multiple of 10 from 10 to 10,000, move the decimal point one place to the left, look up this number in the N column, and find the square root in the $\sqrt{10N}$ column. For example, to get the square root of 8,670, look up 867 in the N column and find 93.1 + in the $\sqrt{10N}$ column.
3. When a problem calls for the square root of a number not given in the N column, it may be possible to find that number in the N^2 column. If the number is located in the N^2 column, its square root is given in the N column. Thus, to obtain the square root of 1,225, find this number under N^2 and read the square root, 35, to the left in the N column.

The square root of other numbers may also be read from the table in any of these methods by observing the rule that moving the decimal point *two* places to the left or right in the number moves it *one* place in the square root. As an example, the number 123,201 is given in the N^2 column of the table. Then,

The square root of 123,201. = 351.

The square root of 1,232.01 = 35.1

The square root of 12.3201 = 3.51

The square root of any number not shown in the table may be estimated by interpolating between values which are included. For exam-

ple, the square root of 65.12 must be between the square root of 65 and the square root of 66. Since 65.12 stands at a point .12 of the way from 65 to 66, its square root should be approximately .12 of the way from the square root of 65 to the square root of 66. The following procedure is used:

Number	Root
66.....	8.124
65.12.....	?
65.....	8.062
Difference.....	.062

$$\sqrt{65.12} = 8.062 + .12(.062) = 8.069 +$$

More detailed values of squares roots may be obtained without interpolation by the use of *Barlow's Tables*, published by the Chemical Publishing Co., Inc., 234 King Street, Brooklyn, New York, which gives the squares, cubes, square roots, cube roots, and reciprocals of all integer numbers up to 12,500.

THE USE OF RECIPROCAL

Multiplication and division can often be facilitated by the use of reciprocals.* Instead of multiplying, one can divide one number by the reciprocal of the second, if the reciprocal is a simple number. For example:

$$1,582 \times 25 = \frac{158,200}{4} = 39,550$$

$$220 \times 50 = \frac{22,000}{2} = 11,000$$

$$17,228 \times 125 = \frac{17,228,000}{8} = 2,153,500$$

Similarly, instead of dividing, it may be easier to multiply the numerator by the reciprocal of the denominator. Thus,

$$5,725 \div 25 = 5,725 \times .04 = 57.25 \times 4 = 229$$

$$280,400 \div 50 = 2,804 \times 2 = 5,608$$

$$245,925 \div 125 = 245.925 \times 8 = 1,967.4$$

This short cut is particularly useful in computing a series of percents on a common base, such as percents of various asset accounts to total assets in a balance sheet. Simply place the reciprocal of the base (e.g., total assets) in a calculating machine, and *multiply* by each of the other items in turn, without clearing the machine. Reciprocals may be found in the last column of the following table.

* The reciprocal of a number is defined as unity divided by the number; i.e., the reciprocal of 5 is $1 \div 5 = .2$. The reciprocal of .25 is $1 \div .25 = 4$.

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000*

N	N²	√N	√10N	1/N	N	N²	√N	√10N	1/N .0
1	1	1.000 000	3.162 278	.10000000	50	2 500	7.071 068	22.36068	2000000
2	4	1.414 214	4.472 136	.50000000	51	2 601	7.141 428	22.58318	1960784
3	9	1.732 051	5.477 226	.33333333	52	2 704	7.211 103	22.80351	1923077
4	16	2.000 000	6.324 555	.25000000	53	2 809	7.280 110	23.02173	1886792
5	25	2.236 068	7.071 068	.20000000	54	2 916	7.348 469	23.23790	1851852
6	36	2.449 490	7.745 967	.16666667	55	3 025	7.416 198	23.45208	1818182
7	49	2.645 751	8.366 600	.14285714	56	3 136	7.483 315	23.66432	1785714
8	64	2.828 427	8.944 272	.12500000	57	3 249	7.549 834	23.87467	1754386
9	81	3.000 000	9.486 833	.11111111	58	3 364	7.615 773	24.08319	1724138
10	100	3.162 278	10.00000	.10000000	59	3 481	7.681 146	24.28992	1694915
11	121	3.316 625	10.48809	.09090909	60	3 600	7.745 967	24.49490	1666667
12	144	3.464 102	10.95445	.08333333	61	3 721	7.810 250	24.69818	1639344
13	169	3.605 551	11.40175	.07692308	62	3 844	7.874 008	24.89980	1612903
14	196	3.741 657	11.83216	.07142857	63	3 969	7.937 254	25.09980	1587302
15	225	3.872 983	12.24745	.06666667	64	4 096	8.000 000	25.29822	1562500
16	256	4.000 000	12.64911	.06250000	65	4 225	8.062 258	25.49510	1538462
17	289	4.123 106	13.03840	.05882353	66	4 356	8.124 038	25.69047	1515152
18	324	4.242 641	13.41641	.05555556	67	4 489	8.185 353	25.88436	1492537
19	361	4.358 899	13.78405	.05263158	68	4 624	8.246 211	26.07681	1470588
20	400	4.472 136	14.14214	.05000000	69	4 761	8.306 624	26.26785	1449275
21	441	4.582 576	14.49138	.04761905	70	4 900	8.366 600	26.45751	1428571
22	484	4.690 416	14.83240	.04545455	71	5 041	8.426 150	26.64583	1408451
23	529	4.795 832	15.16575	.04347826	72	5 184	8.485 281	26.83282	1388889
24	576	4.898 979	15.49193	.04166667	73	5 329	8.544 004	27.01851	1369863
25	625	5.000 000	15.81139	.04000000	74	5 476	8.602 325	27.20294	1351351
26	676	5.099 020	16.12452	.03846154	75	5 625	8.660 254	27.38613	1333333
27	729	5.196 152	16.43168	.03703704	76	5 776	8.717 798	27.56810	1315789
28	784	5.291 503	16.73320	.03571429	77	5 929	8.774 964	27.74887	1298701
29	841	5.385 165	17.02939	.03448276	78	6 084	8.831 761	27.92848	1282051
30	900	5.477 226	17.32051	.03333333	79	6 241	8.888 194	28.10694	1265823
31	961	5.567 764	17.60682	.03225806	80	6 400	8.944 272	28.28427	1250000
32	1 024	5.656 854	17.88854	.03125000	81	6 561	9.000 000	28.46050	1234568
33	1 089	5.744 563	18.16590	.03030303	82	6 724	9.055 385	28.63564	1219512
34	1 156	5.830 952	18.43909	.02941176	83	6 889	9.110 434	28.80972	1204819
35	1 225	5.916 080	18.70829	.02857143	84	7 056	9.165 151	28.98275	1190476
36	1 296	6.000 000	18.97367	.02777778	85	7 225	9.219 544	29.15476	1176471
37	1 369	6.082 763	19.23538	.02702703	86	7 396	9.273 618	29.32576	1162791
38	1 444	6.164 414	19.49359	.02631579	87	7 569	9.327 379	29.49576	1149425
39	1 521	6.244 998	19.74842	.02564103	88	7 744	9.380 832	29.66479	1136364
40	1 600	6.324 555	20.00000	.02500000	89	7 921	9.433 981	29.83287	1123596
41	1 681	6.403 124	20.24846	.02439024	90	8 100	9.486 833	30.00000	1111111
42	1 764	6.480 741	20.49390	.02380952	91	8 281	9.539 392	30.16621	1098901
43	1 849	6.557 439	20.73644	.02325581	92	8 464	9.591 663	30.33150	1086957
44	1 936	6.633 250	20.97618	.02272727	93	8 649	9.643 651	30.49590	1075269
45	2 025	6.708 204	21.21320	.02222222	94	8 836	9.695 360	30.65942	1063830
46	2 116	6.782 330	21.44761	.02173913	95	9 025	9.746 794	30.82207	1052632
47	2 209	6.855 655	21.67948	.02127660	96	9 216	9.797 959	30.98387	1041667
48	2 304	6.928 203	21.90890	.02083333	97	9 409	9.848 858	31.14482	1030928
49	2 401	7.000 000	22.13594	.02040816	98	9 604	9.899 495	31.30495	1020408
50	2 500	7.071 068	22.36068	.02000000	99	9 801	9.949 874	31.46427	1010101
					100	10 000	10.00000	31.62278	1000000

* From Frederick E. Croxton and Dudley J. Cowden, *Practical Business Statistics*, © 1948, pp. 524-33. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.0	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
100	10 000	10.00000	31.62278	10000000	150	22 500	12.24745	38.72983	6666667
101	10 201	10.04988	31.78050	09900990	151	22 801	12.28821	38.85872	6622517
102	10 404	10.09950	31.93744	09803922	152	23 104	12.32883	38.98718	6578947
103	10 609	10.14889	32.09361	09708738	153	23 409	12.36932	39.11521	6535948
104	10 816	10.19804	32.24903	09615385	154	23 716	12.40967	39.24283	6493506
105	11 025	10.24695	32.40370	09523810	155	24 025	12.44990	39.37004	6451613
106	11 236	10.29563	32.55764	09433962	156	24 336	12.49000	39.49684	6410256
107	11 449	10.34408	32.71085	09345794	157	24 649	12.52996	39.62323	6369427
108	11 664	10.39230	32.86335	09259259	158	24 964	12.56981	39.74921	6329114
109	11 881	10.44031	33.01515	09174312	159	25 281	12.60952	39.87480	6289308
110	12 100	10.48809	33.16625	09090909	160	25 600	12.64911	40.00000	6250000
111	12 321	10.53565	33.31666	09009009	161	25 921	12.68858	40.12481	6211180
112	12 544	10.58301	33.46640	08928571	162	26 244	12.72792	40.24922	6172840
113	12 769	10.63015	33.61547	08849558	163	26 569	12.76715	40.37326	6134969
114	12 996	10.67708	33.76389	08771930	164	26 896	12.80625	40.49691	6097561
115	13 225	10.72381	33.91165	08695652	165	27 225	12.84523	40.62019	6060606
116	13 456	10.77033	34.05877	08620690	166	27 556	12.88410	40.74310	6024096
117	13 689	10.81665	34.20526	08547009	167	27 889	12.92285	40.86563	5988024
118	13 924	10.86278	34.35113	08474576	168	28 224	12.96148	40.98780	5952381
119	14 161	10.90871	34.49638	08403361	169	28 561	13.00000	41.10961	5917160
120	14 400	10.95445	34.64102	08333333	170	28 900	13.03840	41.23106	5882353
121	14 641	11.00000	34.78505	08264463	171	29 241	13.07670	41.35215	5847953
122	14 884	11.04536	34.92850	08196721	172	29 584	13.11488	41.47288	5813953
123	15 129	11.09054	35.07136	08130081	173	29 929	13.15295	41.59327	5780347
124	15 376	11.13553	35.21363	08064516	174	30 276	13.19091	41.71331	5747126
125	15 625	11.18034	35.35534	08000000	175	30 625	13.22876	41.83300	5714286
126	15 876	11.22497	35.49648	07936508	176	30 976	13.26650	41.95235	5681818
127	16 129	11.26943	35.63706	07874016	177	31 329	13.30413	42.07137	5649718
128	16 384	11.31371	35.77709	07812500	178	31 684	13.34166	42.19005	5617978
129	16 641	11.35782	35.91657	07751938	179	32 041	13.37909	42.30839	5586592
130	16 900	11.40175	36.05551	07692308	180	32 400	13.41641	42.42641	5555556
131	17 161	11.44552	36.19392	07633588	181	32 761	13.45362	42.54409	5524862
132	17 424	11.48913	36.33180	07575758	182	33 124	13.49074	42.66146	5494505
133	17 689	11.53256	36.46917	07518797	183	33 489	13.52775	42.77850	5464481
134	17 956	11.57584	36.60601	07462687	184	33 856	13.56466	42.89522	5434783
135	18 225	11.61895	36.74235	07407407	185	34 225	13.60147	43.01163	5405405
136	18 496	11.66190	36.87818	07352941	186	34 596	13.63818	43.12772	5376344
137	18 769	11.70470	37.01351	07299270	187	34 969	13.67479	43.24350	5347594
138	19 044	11.74734	37.14835	07246377	188	35 344	13.71131	43.35897	5319149
139	19 321	11.78983	37.28270	07194245	189	35 721	13.74773	43.47413	5291005
140	19 600	11.83216	37.41657	07142857	190	36 100	13.78405	43.58899	5263158
141	19 881	11.87434	37.54997	07092199	191	36 481	13.82027	43.70355	5235602
142	20 164	11.91638	37.68289	07042254	192	36 864	13.85641	43.81780	5208333
143	20 449	11.95826	37.81534	06993007	193	37 249	13.89244	43.93177	5181347
144	20 736	12.00000	37.94733	06944444	194	37 636	13.92839	44.04543	5154639
145	21 025	12.04159	38.07887	06896552	195	38 025	13.96424	44.15880	5128205
146	21 316	12.08305	38.20995	06849315	196	38 416	14.00000	44.27189	5102041
147	21 609	12.12436	38.34058	06802721	197	38 809	14.03567	44.38468	5076142
148	21 904	12.16553	38.47077	06756757	198	39 204	14.07125	44.49719	5050505
149	22 201	12.20656	38.60052	06711409	199	39 601	14.10674	44.60942	5025126
150	22 500	12.24745	38.72983	06666667	200	40 000	14.14214	44.72136	5000000

SQUARES, SQUARE ROOTS, AND RECIPROALS 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
200	40 000	14.14214	44.72136	5000000	250	62 500	15.81139	50.00000	4000000
201	40 401	14.17745	44.83302	4975124	251	63 001	15.84298	50.09990	3984064
202	40 804	14.21287	44.94441	4950495	252	63 504	15.87451	50.19960	3968254
203	41 209	14.24781	45.05552	4926108	253	64 009	15.90597	50.29911	3952569
204	41 616	14.28286	45.16636	4901961	254	64 516	15.93738	50.39841	3937008
205	42 025	14.31782	45.27693	4878049	255	65 025	15.96872	50.49752	3921569
206	42 436	14.35270	45.38722	4854369	256	65 536	16.00000	50.59644	3906250
207	42 849	14.38749	45.49725	4830918	257	66 049	16.03122	50.69517	3891051
208	43 264	14.42221	45.60702	4807692	258	66 564	16.06238	50.79370	3875969
209	43 681	14.45683	45.71652	4784689	259	67 081	16.09348	50.89204	3861004
210	44 100	14.49138	45.82576	4761905	260	67 600	16.12452	50.99020	3846154
211	44 521	14.52584	45.93474	4739336	261	68 121	16.15549	51.08816	3831418
212	44 944	14.56022	46.04346	4716981	262	68 644	16.18641	51.18594	3816794
213	45 369	14.59452	46.15192	4694836	263	69 169	16.21727	51.28353	3802281
214	45 796	14.62874	46.26013	4672897	264	69 696	16.24808	51.38093	3787879
215	46 225	14.66288	46.36809	4651163	265	70 225	16.27882	51.47815	3773585
216	46 656	14.69694	46.47580	4629630	266	70 756	16.30951	51.57519	3759398
217	47 089	14.73092	46.58326	4608295	267	71 289	16.34013	51.67204	3745318
218	47 524	14.76482	46.69047	4587158	268	71 824	16.37071	51.76872	3731343
219	47 961	14.79865	46.79744	4566210	269	72 361	16.40122	51.86521	3717472
220	48 400	14.83240	46.90416	4545455	270	72 900	16.43168	51.96152	3703704
221	48 841	14.86607	47.01064	4524887	271	73 441	16.46208	52.05766	3690037
222	49 284	14.89966	47.11688	4504505	272	73 984	16.49242	52.15362	3676471
223	49 729	14.93318	47.22288	4484305	273	74 529	16.52271	52.24940	3663004
224	50 176	14.96663	47.32864	4464286	274	75 076	16.55295	52.34501	3649635
225	50 625	15.00000	47.43416	4444444	275	75 625	16.58312	52.44044	3636364
226	51 076	15.03330	47.53946	4424779	276	76 176	16.61325	52.53570	3623188
227	51 529	15.06652	47.64452	4405286	277	76 729	16.64332	52.63079	3610108
228	51 984	15.09967	47.74935	4385965	278	77 284	16.67333	52.72571	3597122
229	52 441	15.13275	47.85394	4366812	279	77 841	16.70329	52.82045	3584229
230	52 900	15.16575	47.95832	4347826	280	78 400	16.73320	52.91503	3571429
231	53 361	15.19868	48.06246	4329004	281	78 961	16.76305	53.00943	3558719
232	53 824	15.23155	48.16638	4310345	282	79 524	16.79286	53.10367	3546099
233	54 289	15.26434	48.27007	4291845	283	80 089	16.82260	53.19774	3533569
234	54 756	15.29706	48.37355	4273504	284	80 656	16.85230	53.29165	3521127
235	55 225	15.32971	48.47680	4255319	285	81 225	16.88194	53.38539	3508772
236	55 696	15.36229	48.57983	4237288	286	81 796	16.91153	53.47897	3496503
237	56 169	15.39480	48.68265	4219409	287	82 369	16.94107	53.57238	3484321
238	56 644	15.42725	48.78524	4201681	288	82 944	16.97056	53.66563	3472222
239	57 121	15.45962	48.88763	4184100	289	83 521	17.00000	53.75872	3460208
240	57 600	15.49193	48.98979	4166667	290	84 100	17.02939	53.85165	3448276
241	58 081	15.52417	49.09175	4149378	291	84 681	17.05872	53.94442	3436426
242	58 564	15.55635	49.19350	4132231	292	85 264	17.08801	54.03702	3424658
243	59 049	15.58846	49.29503	4115226	293	85 849	17.11724	54.12947	3412969
244	59 536	15.62050	49.39636	4098361	294	86 436	17.14643	54.22177	3401361
245	60 025	15.65248	49.49747	4081633	295	87 025	17.17556	54.31390	3389831
246	60 516	15.68439	49.59839	4065041	296	87 616	17.20465	54.40588	3378378
247	61 009	15.71623	49.69909	4048583	297	88 209	17.23369	54.49771	3367003
248	61 504	15.74802	49.79960	4032258	298	88 804	17.26268	54.58938	3355705
249	62 001	15.77973	49.89990	4016064	299	89 401	17.29162	54.68089	3344482
250	62 500	15.81139	50.00000	4000000	300	90 000	17.32051	54.77226	3333333

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
300	90 000	17.32051	54.77226	3333333	350	122 500	18.70829	59.16080	2857143
301	90 601	17.34935	54.86347	3322259	351	123 201	18.73499	59.24525	2849003
302	91 204	17.37815	54.95453	3311258	352	123 904	18.76166	59.32959	2840909
303	91 809	17.40690	55.04544	3300330	353	124 609	18.78829	59.41380	2832861
304	92 416	17.43563	55.13620	3289474	354	125 316	18.81489	59.49790	2824859
305	93 025	17.46425	55.22681	3278689	355	126 025	18.84144	59.58188	2816901
306	93 636	17.49286	55.31727	3267974	356	126 736	18.86796	59.66574	2808989
307	94 249	17.52142	55.40758	3257329	357	127 449	18.89444	59.74948	2801120
308	94 864	17.54993	55.49775	3246753	358	128 164	18.92089	59.83310	2793296
309	95 481	17.57840	55.58777	3236246	359	128 881	18.94730	59.91661	2785515
310	96 100	17.60682	55.67764	3225806	360	129 600	18.97367	60.00000	2777778
311	96 721	17.63519	55.76737	3215434	361	130 321	19.00000	60.08328	2770083
312	97 344	17.66352	55.85696	3205128	362	131 044	19.02630	60.16644	2762431
313	97 969	17.69181	55.94640	3194888	363	131 769	19.05256	60.24948	2754821
314	98 596	17.72005	56.03570	3184713	364	132 496	19.07878	60.33241	2747253
315	99 225	17.74824	56.12486	3174603	365	133 225	19.10497	60.41523	2739726
316	99 856	17.77639	56.21388	3164557	366	133 956	19.13113	60.49793	2732240
317	100 489	17.80449	56.30275	3154574	367	134 689	19.15724	60.58052	2724796
318	101 124	17.83255	56.39149	3144654	368	135 424	19.18333	60.66300	2717391
319	101 761	17.86057	56.48008	3134796	369	136 161	19.20937	60.74537	2710027
320	102 400	17.88854	56.56854	3125000	370	136 900	19.23538	60.82763	2702703
321	103 041	17.91647	56.65686	3115265	371	137 641	19.26136	60.90977	2695418
322	103 684	17.94436	56.74504	3105590	372	138 384	19.28730	60.99180	2688172
323	104 329	17.97220	56.83309	3095975	373	139 129	19.31321	61.07373	2680965
324	104 976	18.00000	56.92100	3086420	374	139 876	19.33908	61.15554	2673797
325	105 625	18.02776	57.00877	3076923	375	140 625	19.36492	61.23724	2666667
326	106 276	18.05547	57.09641	3067485	376	141 376	19.39072	61.31884	2659574
327	106 929	18.08314	57.18391	3058104	377	142 129	19.41649	61.40033	2652520
328	107 584	18.11077	57.27128	3048780	378	142 884	19.44222	61.48170	2645503
329	108 241	18.13836	57.35852	3039514	379	143 641	19.46792	61.56298	2638522
330	108 900	18.16590	57.44563	3030303	380	144 400	19.49359	61.64414	2631579
331	109 561	18.19341	57.53260	3021148	381	145 161	19.51922	61.72520	2624672
332	110 224	18.22087	57.61944	3012048	382	145 924	19.54483	61.80615	2617801
333	110 889	18.24829	57.70615	3003003	383	146 689	19.57039	61.88699	2610966
334	111 556	18.27567	57.79273	2994012	384	147 456	19.59592	61.96773	2604167
335	112 225	18.30301	57.87918	2985075	385	148 225	19.62142	62.04837	2597403
336	112 896	18.33030	57.96551	2976190	386	148 996	19.64688	62.12890	2590674
337	113 569	18.35756	58.05170	2967359	387	149 769	19.67232	62.20932	2583979
338	114 244	18.38478	58.13777	2958580	388	150 544	19.69772	62.28965	2577320
339	114 921	18.41195	58.22371	2949853	389	151 321	19.72308	62.36986	2570694
340	115 600	18.43909	58.30952	2941176	390	152 100	19.74842	62.44998	2564103
341	116 281	18.46619	58.39521	2932551	391	152 881	19.77372	62.52999	2557545
342	116 964	18.49324	58.48077	2923977	392	153 664	19.79899	62.60990	2551020
343	117 649	18.52026	58.56620	2915452	393	154 449	19.82423	62.68971	2544529
344	118 336	18.54724	58.65151	2906977	394	155 236	19.84943	62.76942	2538071
345	119 025	18.57418	58.73670	2898551	395	156 025	19.87461	62.84903	2531646
346	119 716	18.60108	58.82176	2890173	396	156 816	19.89975	62.92853	2525253
347	120 409	18.62794	58.90671	2881844	397	157 609	19.92486	63.00794	2518892
348	121 104	18.65476	58.99152	2873563	398	158 404	19.94994	63.08724	2512563
349	121 801	18.68154	59.07622	2865330	399	159 201	19.97498	63.16645	2506266
350	122 500	18.70829	59.16080	2857143	400	160 000	20.00000	63.24555	2500000

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
400	160 000	20.00000	63.24555	2500000	450	202 500	21.21320	67.08204	2222222
401	160 801	20.02498	63.32456	2493766	451	203 401	21.23676	67.15653	2217295
402	161 604	20.04994	63.40347	2487562	452	204 304	21.26029	67.23095	2212389
403	162 409	20.07486	63.48228	2481390	453	205 209	21.28380	67.30527	2207506
404	163 216	20.09975	63.56099	2475248	454	206 116	21.30728	67.37952	2202643
405	164 025	20.12461	63.63961	2469136	455	207 025	21.33073	67.45369	2197802
406	164 836	20.14944	63.71813	2463054	456	207 936	21.35416	67.52777	2192982
407	165 649	20.17424	63.79655	2457002	457	208 849	21.37756	67.60178	2188184
408	166 464	20.19901	63.87488	2450980	458	209 764	21.40093	67.67570	2183406
409	167 281	20.22375	63.95311	2444988	459	210 681	21.42429	67.74954	2178649
410	168 100	20.24846	64.03124	2439024	460	211 600	21.44761	67.82330	2173913
411	168 921	20.27313	64.10928	2433090	461	212 521	21.47091	67.89698	2169197
412	169 744	20.29778	64.18723	2427184	462	213 444	21.49419	67.97058	2164502
413	170 569	20.32240	64.26508	2421308	463	214 369	21.51743	68.04410	2159827
414	171 396	20.34699	64.34283	2415459	464	215 296	21.54066	68.11755	2155172
415	172 225	20.37155	64.42049	2409639	465	216 225	21.56386	68.19091	2150538
416	173 056	20.39608	64.49806	2403846	466	217 156	21.58703	68.26419	2145923
417	173 889	20.42058	64.57554	2398082	467	218 089	21.61018	68.33740	2141328
418	174 724	20.44505	64.65292	2392344	468	219 024	21.63331	68.41053	2136752
419	175 561	20.46949	64.73021	2386635	469	219 961	21.65641	68.48357	2132196
420	176 400	20.49390	64.80741	2380952	470	220 900	21.67948	68.55655	2127660
421	177 241	20.51828	64.88451	2375297	471	221 841	21.70253	68.62944	2123142
422	178 084	20.54264	64.96153	2369668	472	222 784	21.72556	68.70226	2118644
423	178 929	20.56696	65.03845	2364066	473	223 729	21.74856	68.77500	2114165
424	179 776	20.59126	65.11528	2358491	474	224 676	21.77154	68.84766	2109705
425	180 625	20.61553	65.19202	2352941	475	225 625	21.79449	68.92024	2105263
426	181 476	20.63977	65.26868	2347418	476	226 576	21.81742	68.99275	2100840
427	182 329	20.66398	65.34524	2341920	477	227 529	21.84033	69.06519	2096436
428	183 184	20.68816	65.42171	2336449	478	228 484	21.86321	69.13754	2092050
429	184 041	20.71232	65.49809	2331002	479	229 441	21.88607	69.20983	2087683
430	184 900	20.73644	65.57439	2325581	480	230 400	21.90890	69.28203	2083333
431	185 761	20.76054	65.65059	2320186	481	231 361	21.93171	69.35416	2079002
432	186 624	20.78461	65.72671	2314815	482	232 324	21.95450	69.42622	2074689
433	187 489	20.80865	65.80274	2309469	483	233 289	21.97726	69.49820	2070393
434	188 356	20.83267	65.87868	2304147	484	234 256	22.00000	69.57011	2066116
435	189 225	20.85665	65.95453	2298851	485	235 225	22.02272	69.64194	2061856
436	190 096	20.88061	66.03030	2293578	486	236 196	22.04541	69.71370	2057613
437	190 969	20.90454	66.10598	2288330	487	237 169	22.06808	69.78539	2053388
438	191 844	20.92845	66.18157	2283105	488	238 144	22.09072	69.85700	2049180
439	192 721	20.95233	66.25708	2277904	489	239 121	22.11334	69.92853	2044990
440	193 600	20.97618	66.33250	2272727	490	240 100	22.13594	70.00000	2040816
441	194 481	21.00000	66.40783	2267574	491	241 081	22.15852	70.07139	2036660
442	195 364	21.02380	66.48308	2262443	492	242 064	22.18107	70.14271	2032520
443	196 249	21.04757	66.55825	2257336	493	243 049	22.20360	70.21396	2028398
444	197 136	21.07131	66.63332	2252252	494	244 036	22.22611	70.28513	2024291
445	198 025	21.09502	66.70832	2247191	495	245 025	22.24860	70.35624	2020202
446	198 916	21.11871	66.78323	2242152	496	246 016	22.27106	70.42727	2016129
447	199 809	21.14237	66.85806	2237136	497	247 009	22.29350	70.49823	2012072
448	200 704	21.16601	66.93280	2232143	498	248 004	22.31591	70.56912	2008032
449	201 601	21.18962	67.00746	2227171	499	249 001	22.33831	70.63993	2004008
450	202 500	21.21320	67.08204	2222222	500	250 000	22.36068	70.71068	2000000

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
500	250 000	22.36068	70.71068	2000000	550	302 500	23.45208	74.16198	1818182
501	251 001	22.38303	70.78135	1996008	551	303 601	23.47339	74.22937	1814882
502	252 004	22.40536	70.85196	1992032	552	304 704	23.49468	74.29670	1811594
503	253 009	22.42766	70.92249	1988072	553	305 809	23.51595	74.36397	1808318
504	254 016	22.44994	70.99296	1984127	554	306 916	23.53720	74.43118	1805054
505	255 025	22.47221	71.06335	1980198	555	308 025	23.55844	74.49832	1801802
506	256 036	22.49444	71.13368	1976285	556	309 136	23.57965	74.56541	1798561
507	257 049	22.51666	71.20393	1972387	557	310 249	23.60085	74.63243	1795332
508	258 064	22.53886	71.27412	1968504	558	311 364	23.62202	74.69940	1792115
509	259 081	22.56103	71.34424	1964637	559	312 481	23.64318	74.76630	1788909
510	260 100	22.58318	71.41428	1960784	560	313 600	23.66432	74.83315	1785714
511	261 121	22.60531	71.48426	1956947	561	314 721	23.68544	74.89993	1782531
512	262 144	22.62742	71.55418	1953125	562	315 844	23.70654	74.96666	1779359
513	263 169	22.64950	71.62402	1949318	563	316 969	23.72762	75.03333	1776199
514	264 196	22.67157	71.69379	1945525	564	318 096	23.74868	75.09993	1773050
515	265 225	22.69361	71.76350	1941748	565	319 225	23.76973	75.16648	1769912
516	266 256	22.71563	71.83314	1937984	566	320 356	23.79075	75.23297	1766784
517	267 289	22.73763	71.90271	1934236	567	321 489	23.81176	75.29940	1763668
518	268 324	22.75961	71.97222	1930502	568	322 624	23.83275	75.36577	1760563
519	269 361	22.78157	72.04165	1926782	569	323 761	23.85372	75.43209	1757469
520	270 400	22.80351	72.11103	1923077	570	324 900	23.87467	75.49834	1754386
521	271 441	22.82542	72.18033	1919386	571	326 041	23.89561	75.56454	1751313
522	272 484	22.84732	72.24957	1915709	572	327 184	23.91652	75.63068	1748252
523	273 529	22.86919	72.31874	1912046	573	328 329	23.93742	75.69676	1745201
524	274 576	22.89105	72.38784	1908397	574	329 476	23.95830	75.76279	1742160
525	275 625	22.91288	72.45688	1904762	575	330 625	23.97916	75.82875	1739130
526	276 676	22.93469	72.52586	1901141	576	331 776	24.00000	75.89466	1736111
527	277 729	22.95648	72.59477	1897533	577	332 929	24.02082	75.96052	1733102
528	278 784	22.97825	72.66361	1893939	578	334 084	24.04163	76.02631	1730104
529	279 841	23.00000	72.73239	1890359	579	335 241	24.06242	76.09205	1727116
530	280 900	23.02173	72.80110	1886792	580	336 400	24.08319	76.15773	1724138
531	281 961	23.04344	72.86975	1883239	581	337 561	24.10394	76.22336	1721170
532	283 024	23.06513	72.93833	1879699	582	338 724	24.12468	76.28892	1718213
533	284 089	23.08679	73.00685	1876173	583	339 889	24.14539	76.35444	1715266
534	285 156	23.10844	73.07530	1872659	584	341 056	24.16609	76.41989	1712329
535	286 225	23.13007	73.14369	1869159	585	342 225	24.18677	76.48529	1709402
536	287 296	23.15167	73.21202	1865672	586	343 396	24.20744	76.55064	1706485
537	288 369	23.17326	73.28028	1862197	587	344 569	24.22808	76.61593	1703578
538	289 444	23.19483	73.34848	1858736	588	345 744	24.24871	76.68116	1700680
539	290 521	23.21637	73.41662	1855288	589	346 921	24.26932	76.74634	1697793
540	291 600	23.23790	73.48469	1851852	590	348 100	24.28992	76.81146	1694915
541	292 681	23.25941	73.55270	1848429	591	349 281	24.31049	76.87652	1692047
542	293 764	23.28089	73.62065	1845018	592	350 464	24.33105	76.94154	1689189
543	294 849	23.30236	73.68853	1841621	593	351 649	24.35159	77.00649	1686341
544	295 936	23.32381	73.75636	1838235	594	352 836	24.37212	77.07140	1683502
545	297 025	23.34524	73.82412	1834862	595	354 025	24.39262	77.13624	1680672
546	298 116	23.36664	73.89181	1831502	596	355 216	24.41311	77.20104	1677852
547	299 209	23.38803	73.95945	1828154	597	356 409	24.43358	77.26578	1675042
548	300 304	23.40940	74.02702	1824818	598	357 604	24.45404	77.33046	1672241
549	301 401	23.43075	74.09453	1821494	599	358 801	24.47448	77.39509	1669449
550	302 500	23.45208	74.16198	1818182	600	360 000	24.49490	77.45967	1666667

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N ²	\sqrt{N}	$\sqrt{10N}$	1/N .00	N	N ²	\sqrt{N}	$\sqrt{10N}$	1/N .00
600	360 000	24.49490	77.45967	1666667	650	422 500	25.49510	80.62258	1538462
601	361 201	24.51530	77.52419	1663894	651	423 801	25.51470	80.68457	1536098
602	362 404	24.53569	77.58866	1661130	652	425 104	25.53429	80.74652	1533742
603	363 609	24.55606	77.65307	1658375	653	426 409	25.55386	80.80842	1531394
604	364 816	24.57641	77.71744	1655629	654	427 716	25.57342	80.87027	1529052
605	366 025	24.59675	77.78175	1652893	655	429 025	25.59297	80.93207	1526718
606	367 236	24.61707	77.84600	1650165	656	430 336	25.61250	80.99383	1524390
607	368 449	24.63737	77.91020	1647446	657	431 649	25.63201	81.05554	1522070
608	369 664	24.65766	77.97435	1644737	658	432 964	25.65151	81.11720	1519757
609	370 881	24.67793	78.03845	1642036	659	434 281	25.67100	81.17881	1517451
610	372 100	24.69818	78.10250	1639344	660	435 600	25.69047	81.24038	1515152
611	373 321	24.71841	78.16649	1636661	661	436 921	25.70992	81.30191	1512859
612	374 544	24.73863	78.23043	1633987	662	438 244	25.72936	81.36338	1510574
613	375 769	24.75884	78.29432	1631321	663	439 569	25.74879	81.42481	1508296
614	376 996	24.77902	78.35815	1628664	664	440 896	25.76820	81.48620	1506024
615	378 225	24.79919	78.42194	1626016	665	442 225	25.78759	81.54753	1503759
616	379 456	24.81935	78.48567	1623377	666	443 556	25.80698	81.60882	1501502
617	380 689	24.83948	78.54935	1620746	667	444 889	25.82634	81.67007	1499250
618	381 924	24.85961	78.61298	1618123	668	446 224	25.84570	81.73127	1497006
619	383 161	24.87971	78.67655	1615509	669	447 561	25.86503	81.79242	1494768
620	384 400	24.89980	78.74008	1612903	670	448 900	25.88436	81.85353	1492537
621	385 641	24.91987	78.80355	1610306	671	450 241	25.90367	81.91459	1490313
622	386 884	24.93993	78.86698	1607717	672	451 584	25.92296	81.97561	1488095
623	388 129	24.95997	78.93035	1605136	673	452 929	25.94224	82.03658	1485884
624	389 376	24.97999	78.99367	1602564	674	454 276	25.96151	82.09750	1483680
625	390 625	25.00000	79.05694	1600000	675	455 625	25.98076	82.15838	1481481
626	391 876	25.01999	79.12016	1597444	676	456 976	26.00000	82.21922	1479290
627	393 129	25.03997	79.18333	1594896	677	458 329	26.01922	82.28001	1477105
628	394 384	25.05993	79.24645	1592357	678	459 684	26.03843	82.34076	1474926
629	395 641	25.07987	79.30952	1589825	679	461 041	26.05763	82.40146	1472754
630	396 900	25.09980	79.37254	1587302	680	462 400	26.07681	82.46211	1470588
631	398 161	25.11971	79.43551	1584786	681	463 761	26.09598	82.42272	1468429
632	399 424	25.13961	79.49843	1582278	682	465 124	26.11513	82.58329	1466276
633	400 689	25.15949	79.56130	1579779	683	466 489	26.13427	82.64381	1464129
634	401 956	25.17936	79.62412	1577287	684	467 856	26.15339	82.70429	1461988
635	403 225	25.19921	79.68689	1574803	685	469 225	26.17250	82.76473	1459854
636	404 496	25.21904	79.74961	1572327	686	470 596	26.19160	82.82512	1457726
637	405 769	25.23886	79.81228	1569859	687	471 969	26.21068	82.88546	1455604
638	407 044	25.25866	79.87490	1567398	688	473 344	26.22975	82.94577	1453488
639	408 321	25.27845	79.93748	1564945	689	474 721	26.24881	83.00602	1451379
640	409 600	25.29822	80.00000	1562500	690	476 100	26.26785	83.06624	1449275
641	410 881	25.31798	80.06248	1560062	691	477 481	26.28688	83.12641	1447178
642	412 164	25.33772	80.12490	1557632	692	478 864	26.30589	83.18654	1445087
643	413 449	25.35744	80.18728	1555210	693	480 249	26.32489	83.24662	1443001
644	414 736	25.37716	80.24961	1552795	694	481 636	26.34388	83.30666	1440922
645	416 025	25.39685	80.31189	1550388	695	483 025	26.36285	83.36666	1438849
646	417 316	25.41653	80.37413	1547988	696	484 416	26.38181	83.42661	1436782
647	418 609	25.43619	80.43631	1545595	697	485 809	26.40076	83.48653	1434720
648	419 904	25.45584	80.49845	1543210	698	487 204	26.41969	83.54639	1432665
649	421 201	25.47548	80.56054	1540832	699	488 601	26.43861	83.60622	1430615
650	422 500	25.49510	80.62258	1538462	700	490 000	26.45751	83.66600	1428571

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N ²	\sqrt{N}	$\sqrt{10N}$	1/N .00	N	N ²	\sqrt{N}	$\sqrt{10N}$	1/N .00
700	490 000	26.45751	83.66600	1428571	750	562 500	27.38613	86.60254	1333333
701	491 401	26.47640	83.72574	1426534	751	564 001	27.40438	86.66026	1331558
702	492 804	26.49528	83.78544	1424501	752	565 504	27.42262	86.71793	1329787
703	494 209	26.51415	83.84510	1422475	753	567 009	27.44085	86.77557	1328021
704	495 616	26.53300	83.90471	1420455	754	568 516	27.45906	86.83317	1326260
705	497 025	26.55184	83.96428	1418440	755	570 025	27.47726	86.89074	1324503
706	498 436	26.57066	84.02381	1416431	756	571 536	27.49545	86.94826	1322751
707	499 849	26.58947	84.08329	1414427	757	573 049	27.51363	87.00575	1321004
708	501 264	26.60827	84.14274	1412429	758	574 564	27.53180	87.06320	1319261
709	502 681	26.62705	84.20214	1410437	759	576 081	27.54995	87.12061	1317523
710	504 100	26.64583	84.26150	1408451	760	577 600	27.56810	87.17798	1315789
711	505 521	26.66458	84.32082	1406470	761	579 121	27.58623	87.23531	1314060
712	506 944	26.68333	84.38009	1404494	762	580 644	27.60435	87.29261	1312336
713	508 369	26.70206	84.43933	1402525	763	582 169	27.62245	87.34987	1310616
714	509 796	26.72078	84.49852	1400560	764	583 696	27.64055	87.40709	1308901
715	511 225	26.73948	84.55767	1398601	765	585 225	27.65863	87.46428	1307190
716	512 656	26.75818	84.61678	1396648	766	586 756	27.67671	87.52143	1305483
717	514 089	26.77686	84.67585	1394700	767	588 289	27.69476	87.57854	1303781
718	515 524	26.79552	84.73488	1392758	768	589 824	27.71281	87.63561	1302083
719	516 961	26.81418	84.79387	1390821	769	591 361	27.73085	87.69265	1300390
720	518 400	26.83282	84.85281	1388889	770	592 900	27.74887	87.74964	1298701
721	519 841	26.85144	84.91172	1386963	771	594 441	27.76689	87.80661	1297017
722	521 284	26.87006	84.97058	1385042	772	595 984	27.78489	87.86353	1295337
723	522 729	26.88866	85.02941	1383126	773	597 529	27.80288	87.92042	1293661
724	524 176	26.90725	85.08819	1381215	774	599 076	27.82086	87.97727	1291990
725	525 625	26.92582	85.14693	1379310	775	600 625	27.83882	88.03408	1290323
726	527 076	26.94439	85.20563	1377410	776	602 176	27.85678	88.09086	1288660
727	528 529	26.96294	85.26429	1375516	777	603 729	27.87472	88.14760	1287001
728	529 984	26.98148	85.32292	1373626	778	605 284	27.89265	88.20431	1285347
729	531 441	27.00000	85.38150	1371742	779	606 841	27.91057	88.26098	1283697
730	532 900	27.01851	85.44004	1369863	780	608 400	27.92848	88.31761	1282051
731	534 361	27.03701	85.49854	1367989	781	609 961	27.94638	88.37420	1280410
732	535 824	27.05550	85.55700	1366120	782	611 524	27.96426	88.43076	1278772
733	537 289	27.07397	85.61542	1364256	783	613 089	27.98214	88.48729	1277139
734	538 756	27.09243	85.67380	1362398	784	614 656	28.00000	88.54377	1275510
735	540 225	27.11088	85.73214	1360544	785	616 225	28.01785	88.60023	1273885
736	541 696	27.12932	85.79044	1358696	786	617 796	28.03569	88.65664	1272265
737	543 169	27.14774	85.84870	1356852	787	619 369	28.05352	88.71302	1270648
738	544 644	27.16616	85.90693	1355014	788	620 944	28.07134	88.76936	1269036
739	546 121	27.18455	85.96511	1353180	789	622 521	28.08914	88.82567	1267427
740	547 600	27.20294	86.02325	1351351	790	624 100	28.10694	88.88194	1265823
741	549 081	27.22132	86.08136	1349528	791	625 681	28.12472	88.93818	1264223
742	550 564	27.23968	86.13942	1347709	792	627 264	28.14249	88.99438	1262626
743	552 049	27.25803	86.19745	1345895	793	628 849	28.16026	89.05055	1261034
744	553 536	27.27636	86.25543	1344086	794	630 436	28.17801	89.10668	1259446
745	555 025	27.29469	86.31338	1342282	795	632 025	28.19574	89.16277	1257862
746	556 516	27.31300	86.37129	1340483	796	633 616	28.21347	89.21883	1256281
747	558 009	27.33130	86.42916	1338688	797	635 209	28.23119	89.27486	1254705
748	559 504	27.34959	86.48699	1336898	798	636 804	28.24889	89.33085	1253133
749	561 001	27.36786	86.54479	1335113	799	638 401	28.26659	89.38680	1251564
750	562 500	27.38613	86.60254	1333333	800	640 000	28.28427	89.44272	1250000

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
800	640 000	28.28427	89.44272	1250000	850	722 500	29.15476	92.19544	1176471
801	641 601	28.30194	89.49860	1248439	851	724 201	29.17190	92.24966	1175088
802	643 204	28.31960	89.55445	1246883	852	725 904	29.18904	92.30385	1173709
803	644 809	28.33725	89.61027	1245330	853	727 609	29.20616	92.35800	1172333
804	646 416	28.35489	89.66605	1243781	854	729 316	29.22328	92.41212	1170960
805	648 025	28.37252	89.72179	1242236	855	731 025	29.24038	92.46621	1169591
806	649 636	28.39014	89.77750	1240695	856	732 736	29.25748	92.52027	1168224
807	651 249	28.40775	89.83318	1239157	857	734 449	29.27456	92.57429	1166861
808	652 864	28.42534	89.88882	1237624	858	736 164	29.29164	92.62829	1165501
809	654 481	28.44293	89.94443	1236094	859	737 881	29.30870	92.68225	1164144
810	656 100	28.46050	90.00000	1234568	860	739 600	29.32576	92.73618	1162791
811	657 721	28.47806	90.05554	1233046	861	741 321	29.34280	92.79009	1161440
812	659 344	28.49561	90.11104	1231527	862	743 044	29.35984	92.84396	1160093
813	660 969	28.51315	90.16651	1230012	863	744 769	29.37686	92.89779	1158749
814	662 596	28.53069	90.22195	1228501	864	746 496	29.39388	92.95160	1157407
815	664 225	28.54820	90.27735	1226994	865	748 225	29.41088	93.00538	1156069
816	665 856	28.56571	90.33272	1225490	866	749 956	29.42788	93.05912	1154734
817	667 489	28.58321	90.38805	1223990	867	751 689	29.44486	93.11283	1153403
818	669 124	28.60070	90.44335	1222494	868	753 424	29.46185	93.16652	1152074
819	670 761	28.61818	90.49862	1221001	869	755 161	29.47881	93.22017	1150748
820	672 400	28.63564	90.55385	1219512	870	756 900	29.49576	93.27379	1149425
821	674 041	28.65310	90.60905	1218027	871	758 641	29.51271	93.32738	1148106
822	675 684	28.67054	90.66422	1216545	872	760 384	29.52965	93.38094	1146789
823	677 329	28.68798	90.71935	1215067	873	762 129	29.54657	93.43447	1145475
824	678 976	28.70540	90.77445	1213592	874	763 876	29.56349	93.48797	1144165
825	680 625	28.72281	90.82951	1212121	875	765 625	29.58040	93.54143	1142857
826	682 276	28.74022	90.88454	1210654	876	767 376	29.59730	93.59487	1141553
827	683 929	28.75761	90.93954	1209190	877	769 129	29.61419	93.64828	1140251
828	685 584	28.77499	90.99451	1207729	878	770 884	29.63106	93.70165	1138952
829	687 241	28.79236	91.04944	1206273	879	772 641	29.64793	93.75500	1137656
830	688 900	28.80972	91.10434	1204819	880	774 400	29.66479	93.80832	1136364
831	690 561	28.82707	91.15920	1203369	881	776 161	29.68164	93.86160	1135074
832	692 224	28.84441	91.21403	1201923	882	777 924	29.69848	93.91486	1133787
833	693 889	28.86174	91.26883	1200480	883	779 689	29.71532	93.96808	1132503
834	695 556	28.87906	91.32360	1199041	884	781 456	29.73214	94.02127	1131222
835	697 225	28.89637	91.37833	1197605	885	783 225	29.74895	94.07444	1129944
836	698 896	28.91366	91.43304	1196172	886	784 996	29.76575	94.12757	1128668
837	700 569	28.93095	91.48770	1194743	887	786 769	29.78255	94.18068	1127396
838	702 244	28.94823	91.54234	1193317	888	788 544	29.79933	94.23375	1126126
839	703 921	28.96550	91.59694	1191895	889	790 321	29.81610	94.28680	1124859
840	705 600	28.98275	91.65151	1190476	890	792 100	29.83287	94.33981	1123596
841	707 281	29.00000	91.70605	1189061	891	793 881	29.84962	94.39280	1122334
842	708 964	29.01724	91.76056	1187648	892	795 664	29.86637	94.44575	1121076
843	710 649	29.03446	91.81503	1186240	893	797 449	29.88311	94.49868	1119821
844	712 336	29.05168	91.86947	1184834	894	799 236	29.89983	94.55157	1118568
845	714 025	29.06888	91.92388	1183432	895	801 025	29.91655	94.60444	1117318
846	715 716	29.08608	91.97826	1182033	896	802 816	29.93326	94.65728	1116071
847	717 409	29.10326	92.03260	1180638	897	804 609	29.94996	94.71008	1114827
848	719 104	29.12044	92.08692	1179245	898	806 404	29.96665	94.76286	1113586
849	720 801	29.13760	92.14120	1177856	899	808 201	29.98333	94.81561	1112347
850	722 500	29.15476	92.19544	1176471	900	810 000	30.00000	94.86833	1111111

SQUARES, SQUARE ROOTS, AND RECIPROCAL 1-1,000 (Continued)

N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00	N	N^2	\sqrt{N}	$\sqrt{10N}$	$1/N$.00
900	810 000	30.00000	94.86833	1111111	950	902 500	30.82207	97.46794	1052632
901	811 801	30.01666	94.92102	1109878	951	904 401	30.83829	97.51923	1051525
902	813 604	30.03331	94.97368	1108647	952	906 304	30.85450	97.57049	1050420
903	815 409	30.04996	95.02631	1107420	953	908 209	30.87070	97.62172	1049318
904	817 216	30.06659	95.07891	1106195	954	910 116	30.88689	97.67292	1048218
905	819 025	30.08322	95.13149	1104972	955	912 025	30.90307	97.72410	1047120
906	820 836	30.09983	95.18403	1103753	956	913 936	30.91925	97.77525	1046025
907	822 649	30.11644	95.23655	1102536	957	915 849	30.93542	97.82638	1044932
908	824 464	30.13304	95.28903	1101322	958	917 764	30.95158	97.87747	1043841
909	826 281	30.14963	95.34149	1100110	959	919 681	30.96773	97.92855	1042753
910	828 100	30.16621	95.39392	1098901	960	921 600	30.98387	97.97959	1041667
911	829 921	30.18278	95.44632	1097695	961	923 521	31.00000	98.03061	1040583
912	831 744	30.19934	95.49869	1096491	962	925 444	31.01612	98.08160	1039501
913	833 569	30.21589	95.55103	1095290	963	927 369	31.03224	98.13256	1038422
914	835 396	30.23243	95.60335	1094092	964	929 296	31.04835	98.18350	1037344
915	837 225	30.24897	95.65563	1092896	965	931 225	31.06445	98.23441	1036269
916	839 056	30.26549	95.70789	1091703	966	933 156	31.08054	98.28530	1035197
917	840 889	30.28201	95.76012	1090513	967	935 089	31.09662	98.33616	1034126
918	842 724	30.29851	95.81232	1089325	968	937 024	31.11270	98.38699	1033058
919	844 561	30.31501	95.86449	1088139	969	938 961	31.12876	98.43780	1031992
920	846 400	30.33150	95.91663	1086957	970	940 900	31.14482	98.48858	1030928
921	848 241	30.34798	95.96874	1085776	971	942 841	31.16087	98.53933	1029866
922	850 084	30.36445	96.02083	1084599	972	944 784	31.17691	98.59006	1028807
923	851 929	30.38092	96.07289	1083424	973	946 729	31.19295	98.64076	1027749
924	853 776	30.39737	96.12492	1082251	974	948 676	31.20897	98.69144	1026694
925	855 625	30.41381	96.17692	1081081	975	950 625	31.22499	98.74209	1025641
926	857 476	30.43025	96.22889	1079914	976	952 576	31.24100	98.79271	1024590
927	859 329	30.44667	96.28084	1078749	977	954 529	31.25700	98.84331	1023541
928	861 184	30.46309	96.33276	1077586	978	956 484	31.27299	98.89388	1022495
929	863 041	30.47950	96.38465	1076426	979	958 441	31.28898	98.94443	1021450
930	864 900	30.49590	96.43651	1075269	980	960 400	31.30495	98.99495	1020408
931	866 761	30.51229	96.48834	1074114	981	962 361	31.32092	99.04544	1019368
932	868 624	30.52868	96.54015	1072961	982	964 324	31.33688	99.09591	1018330
933	870 489	30.54505	96.59193	1071811	983	966 289	31.35283	99.14636	1017294
934	872 356	30.56141	96.64368	1070664	984	968 256	31.36877	99.19677	1016260
935	874 225	30.57777	96.69540	1069519	985	970 225	31.38471	99.24717	1015228
936	876 096	30.59412	96.74709	1068376	986	972 196	31.40064	99.29753	1014199
937	877 969	30.61046	96.79876	1067236	987	974 169	31.41656	99.34787	1013171
938	879 844	30.62679	96.85040	1066098	988	976 144	31.43247	99.39819	1012146
939	881 721	30.64311	96.90201	1064963	989	978 121	31.44837	99.44848	1011122
940	883 600	30.65942	96.95360	1063830	990	980 100	31.46427	99.49874	1010101
941	885 481	30.67572	97.00515	1062699	991	982 081	31.48015	99.54898	1009082
942	887 364	30.69202	97.05668	1061571	992	984 064	31.49603	99.59920	1008065
943	889 249	30.70831	97.10819	1060445	993	986 049	31.51190	99.64939	1007049
944	891 136	30.72458	97.15966	1059322	994	988 036	31.52777	99.69955	1006036
945	893 025	30.74085	97.21111	1058201	995	990 025	31.54362	99.74969	1005025
946	894 916	30.75711	97.26253	1057082	996	992 016	31.55947	99.79980	1004016
947	896 809	30.77337	97.31393	1055966	997	994 009	31.57531	99.84989	1003009
948	898 704	30.78961	97.36529	1054852	998	996 004	31.59114	99.89995	1002004
949	900 601	30.80584	97.41663	1053741	999	998 001	31.60696	99.94999	1001001
950	902 500	30.82207	97.46794	1052632	1000	1 000 000	31.62278	100.00000	1000000

D. AREAS UNDER THE NORMAL CURVE

EACH ENTRY in this table is the proportion of the total area under a normal curve which lies under the segment between the mean and x/σ or u standard deviations from the mean. Example: $x = X - \mu = 31$ and $\sigma = 20$, so $u = x/\sigma = 1.55$. Then the required area is .4394. The area in the tail beyond the point $x = 31$ is then $.5000 - .4394 = .0606$.

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.49903	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993129	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995166	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4996631	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998	.4998
3.5	.4997674	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998409	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4998922	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999277	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.5000	.5000
3.9	.4999519	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000
4.0	.4999683									
4.5	.4999966									
5.0	.4999997133									

SOURCE: Frederick E. Croxton and Dudley J. Cowden, *Practical Business Statistics* (2d ed.; New York: Prentice-Hall, Inc., 1948), p. 511. Reprinted by permission of the publisher.
Through $x/\sigma = 2.99$, from Rugg's *Statistical Methods Applied to Education*, by arrangement with the publishers, Houghton Mifflin Company. A much more detailed table of normal curve areas is given in Federal Works Agency, Work Projects Administration for the City of New York, *Tables of Probability Functions* (New York: National Bureau of Standards, 1942), Vol. II, pp. 2-238. In this appendix values for $x/\sigma = 3.00$ through 5.00 were computed from the latter source.

E. UNIT NORMAL LOSS FUNCTION

THE VALUE $L_N(D)$ is the expected opportunity loss (or EVPI) for a linear loss function with slope one and a unit normal distribution. The value D represents the relative position of the breakeven point.

When using $L_N(D)$ for a general normal distribution, the value D represents the deviation of the breakeven point K from the mean μ , expressed in standard deviation, σ , units. That is $D = \frac{|K - \mu|}{\sigma}$.

D	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.3989	.3940	.3890	.3841	.3793	.3744	.3697	.3649	.3602	.3556
.1	.3509	.3464	.3418	.3373	.3328	.3284	.3240	.3197	.3154	.3111
.2	.3069	.3027	.2986	.2944	.2904	.2863	.2824	.2784	.2745	.2706
.3	.2668	.2630	.2592	.2555	.2518	.2481	.2445	.2409	.2374	.2339
.4	.2304	.2270	.2236	.2203	.2169	.2137	.2104	.2072	.2040	.2009
.5	.1978	.1947	.1917	.1887	.1857	.1828	.1799	.1771	.1742	.1714
.6	.1687	.1659	.1633	.1606	.1580	.1554	.1528	.1503	.1478	.1453
.7	.1429	.1405	.1381	.1358	.1334	.1312	.1289	.1267	.1245	.1223
.8	.1202	.1181	.1160	.1140	.1120	.1100	.1080	.1061	.1042	.1023
.9	.1004	.09860	.09680	.09503	.09328	.09156	.08986	.08819	.08654	.08491
1.0	.08332	.08174	.08019	.07866	.07716	.07568	.07422	.07279	.07138	.06999
1.1	.06862	.06727	.06595	.06465	.06336	.06210	.06086	.05964	.05844	.05726
1.2	.05610	.05496	.05384	.05274	.05165	.05059	.04954	.04851	.04750	.04650
1.3	.04553	.04457	.04363	.04270	.04179	.04090	.04002	.03916	.03831	.03748
1.4	.03667	.03587	.03508	.03431	.03356	.03281	.03208	.03137	.03067	.02998
1.5	.02931	.02865	.02800	.02736	.02674	.02612	.02552	.02494	.02436	.02380
1.6	.02324	.02270	.02217	.02165	.02114	.02064	.02015	.01967	.01920	.01874
1.7	.01829	.01785	.01742	.01699	.01658	.01617	.01578	.01539	.01501	.01464
1.8	.01428	.01392	.01357	.01323	.01290	.01257	.01226	.01195	.01164	.01134
1.9	.01105	.01077	.01049	.01022	.009957	.009698	.009445	.009198	.008957	.008721
2.0	.008491	.008266	.008046	.007832	.007623	.007418	.007219	.007024	.006835	.006649
2.1	.006468	.006292	.006120	.005952	.005788	.005628	.005472	.005320	.005172	.005028
2.2	.004887	.004750	.004616	.004486	.004358	.004235	.004114	.003996	.003882	.003770
2.3	.003662	.003556	.003453	.003352	.003255	.003159	.003067	.002977	.002889	.002804
2.4	.002720	.002640	.002561	.002484	.002410	.002337	.002267	.002199	.002132	.002067
2.5	.002005	.001943	.001883	.001826	.001769	.001715	.001662	.001610	.001560	.001511
3.0	.003822	.003689	.003560	.003436	.003316	.003199	.003087	.002978	.002873	.002771
3.5	.0045848	.0045620	.0045400	.0045188	.0044984	.0044788	.0044599	.0044417	.0044242	.0044073
4.0	.0057145	.0056835	.0056538	.0056253	.0055980	.0055718	.0055468	.0055227	.0054997	.0054777

Reproduced with permission from Robert Schlaifer, *Introduction to Statistics for Business Decisions* (New York: McGraw-Hill 1961) pp. 370-71.

F. BINOMIAL DISTRIBUTION— INDIVIDUAL TERMS

THE TABLE presents individual binomial probabilities for the number of successes, r , in n trials, for selected values of p , the probability of a success on any one trial.

Examples and details in the use of this table for p greater than .50 are given on pages 171 and 172.

The symbol $0+$ indicates a value, positive but less than .0005.

BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS

$$P(r) = {}_nC_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r
2	0	980	960	922	902	884	846	810	774	740	722	706	672	640	608	578	562	490	422	360	302	250	0
	1	020	039	077	095	113	147	180	211	241	255	269	295	320	343	365	375	420	455	480	495	500	1
3	0	970	941	885	857	831	779	729	681	636	614	593	551	512	475	439	422	343	275	216	166	125	0
	1	029	058	111	135	159	203	243	279	311	325	339	363	384	402	416	422	441	444	432	408	375	1
4	0	961	922	849	815	781	716	656	600	547	522	498	452	410	370	334	316	240	179	130	92	63	0
	1	039	075	142	171	199	249	292	327	356	368	379	397	410	418	421	422	412	384	346	299	250	1
5	0	951	904	815	774	734	659	590	528	470	444	418	371	328	289	254	237	168	116	078	050	031	0
	1	048	092	170	204	234	287	328	360	383	392	398	407	410	407	400	396	360	312	259	206	156	1
6	0	941	886	783	735	690	606	531	464	405	377	351	304	262	225	193	178	118	075	047	028	016	0
	1	057	108	196	232	264	316	354	380	395	399	401	400	393	381	365	356	303	244	187	136	094	1
7	0	932	868	751	698	648	558	478	409	348	321	295	249	210	176	146	133	082	049	028	015	008	0
	1	066	124	219	257	290	340	372	390	396	396	393	383	367	347	324	311	247	185	131	087	055	1
8	0	923	851	721	663	610	513	430	360	299	272	248	204	168	137	111	100	058	032	017	008	004	0
	1	075	139	240	279	311	357	383	392	390	385	378	359	336	309	281	267	198	137	090	055	031	1
9	0	914	834	693	630	573	472	387	316	257	232	208	168	134	107	085	075	040	021	010	005	002	0
	1	083	153	260	299	329	370	387	388	377	368	357	331	302	271	240	225	156	100	060	034	018	1
10	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
11	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
12	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
13	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
14	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
15	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
16	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
17	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
18	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
19	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
20	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
21	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
22	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
23	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
24	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
25	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
26	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
27	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
28	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
29	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
30	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188	121	072	040	021	010	1
31	0	904	817	665	599	539	434	349	279	221	197	175	137	107	083	064	056	028	013	006	003	001	0
	1	091	167	277	315	344	378	387	380	360	347	333	302	268	235	203	188</						

BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS (Continued)

$$P(r) = {}_nC_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r
10	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	10
11	0	895	801	638	569	506	400	314	245	190	167	147	113	86	65	49	42	20	09	04	01	0+	0
	1	099	180	293	329	355	382	384	368	341	325	308	272	236	202	170	155	93	52	27	13	05	1
	2	005	018	061	087	113	166	213	251	277	287	293	299	295	284	268	258	200	140	089	051	027	2
	3	0+	001	008	014	022	043	071	103	135	152	168	197	221	241	254	258	257	225	177	126	081	3
	4	0+	0+	001	001	003	008	016	028	044	054	064	086	111	136	160	172	220	243	236	206	161	4
	5	0+	0+	0+	0+	0+	001	002	005	010	013	017	027	039	054	071	080	132	183	221	236	226	5
	6	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	006	010	015	022	027	059	147	193	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	005	006	017	038	070	113	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	004	010	023	046	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	005	013	027	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	11
12	0	886	785	613	540	476	368	282	216	164	142	123	92	69	51	37	32	14	06	02	01	0+	0
	1	107	192	306	341	365	384	377	353	320	301	282	243	206	172	141	127	71	37	17	08	03	1
	2	006	022	070	099	128	183	230	265	286	292	296	294	283	266	244	232	168	109	064	034	016	2
	3	0+	001	010	017	027	053	085	120	155	172	188	215	236	250	257	258	240	195	142	092	054	3
	4	0+	0+	001	002	004	010	021	037	057	068	080	106	133	159	183	194	231	237	213	170	121	4
	5	0+	0+	0+	0+	0+	001	004	008	015	019	025	037	053	072	092	103	158	204	227	222	193	5
	6	0+	0+	0+	0+	0+	0+	0+	001	003	004	005	010	016	024	034	040	079	128	177	212	226	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	006	009	011	029	059	101	149	193	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	002	008	020	042	076	121	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	012	023	054	094	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	007	016	016	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	003	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	12
13	0	878	769	588	513	447	338	254	190	141	121	104	76	55	40	28	24	10	04	01	0+	0+	0
	1	115	204	319	351	371	382	367	336	298	277	257	216	179	145	116	103	54	26	11	04	002	1
	2	007	025	080	111	142	199	245	275	291	294	293	285	268	245	220	206	139	084	045	022	010	2
	3	0+	002	012	021	033	064	100	138	174	190	205	229	246	254	254	252	218	165	111	066	035	3
	4	0+	0+	001	003	005	014	028	047	071	084	098	126	154	179	201	210	234	222	184	135	087	4
	5	0+	0+	0+	0+	001	002	006	012	021	027	033	050	069	091	114	126	180	215	221	199	157	5
	6	0+	0+	0+	0+	0+	001	002	004	006	008	015	023	034	048	056	103	155	197	217	209	177	6
	7	0+	0+	0+	0+	0+	0+	0+	001	001	002	003	006	010	015	019	044	083	131	177	209	177	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	005	014	034	066	109	157	157	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	010	024	050	087	087	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	006	016	035	035	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	010	010	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	12
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	13
14	0	869	754	565	488	421	311	229	167	121	103	87	62	44	31	21	18	07	02	01	0+	0+	0
	1	123	215	329	359	376	379	356	319	276	254	232	191	154	122	95	83	41	18	07	003	001	1
	2	008	029	089	123	156	214	257	283	292	291	287	272	250	223	195	180	113	063	032	014	006	2
	3	0+	002	015	026	040	074	114	154	190	206	219	239	250	252	246	240	194	137	085	046	022	3
	4	0+	0+	002	004	007	018	035	058	085	100	115	144	172	195	214	220	229	202	155	104	061	4
	5	0+	0+	0+	0+	001	003	008	016	028	035	044	063	086	110	135	147	196	218	207	170	122	5
	6	0+	0+	0+	0+	0+	001	003	007	009	012	021	032	047	064	073	126	176	207	209	183	133	6
	7	0+	0+	0+	0+	0+	0+	001	001	002	003	005	009	015	023	028	062	108	157	195	209	177	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	006	008	023	051	092	140	183	133	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	007	018	041	076	122	122	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	014	031	061	061	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	009	022	022	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	006	006	12
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	13
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	14
15	0	860	739	542	463	395	286	206	147	104	87	73	51	35	24	16	13	05	02	0+	0+	0+	0
	1	130	226	339	366	378	373	343	301	254	231	209	168	132	102	77	67	31	13	05	002	0+	1
	2	009	032	099	135	169	227	267	287	290	286	279	258	231	201	171	156	92	048	022	009	003	2
	3	0+	003	018	031	047	086	129	170	204	218	230	245	250	246	234	225	170	111	063	032	014	3
	4	0+	0+	002	005	009	022	043	069	100	116	131	162	188	208	221	225	219	179	127	078	042	4

BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS (Continued)

$$P(r) = {}_n C_r p^r q^{n-r}$$

[illegible]

BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS (*Continued*)

$$P(r) = {}_nC_r p^r q^{n-r}$$

n		r		.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	P		.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r		
19	0	826	681	460	377	309	205	135	088	097	046	036	023	014	009	005	004	001	0+	0+	0+	0+	0+	0+	0+	0+	0		
	1	159	264	364	377	374	339	285	228	176	153	132	096	068	048	033	027	009	003	001	0+	0+	001	0+	0+	0+	1		
	2	014	049	137	179	215	265	285	280	258	243	226	190	154	121	093	080	036	014	005	001	0+	001	0+	0+	0+	2		
	3	001	006	032	053	078	131	180	217	238	243	244	236	218	194	166	152	087	042	017	006	002	006	002	006	002	3		
	4	0+	0+	005	011	020	045	080	118	155	171	186	207	218	219	202	149	091	047	020	007	0+	007	020	007	0+	4		
	5	0+	0+	001	002	004	012	027	048	076	091	106	137	164	185	199	202	192	147	093	050	022	0+	050	022	0+	5		
	6	0+	0+	0+	0+	001	002	007	015	029	037	047	070	095	122	146	157	192	184	145	095	052	0+	095	052	0+	6		
	7	0+	0+	0+	0+	0+	001	004	009	012	017	029	044	064	086	097	153	184	180	144	096	7	0+	184	096	7			
	8	0+	0+	0+	0+	0+	0+	001	002	003	005	009	017	027	041	049	098	149	180	177	144	8	0+	180	177	144	8		
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	001	003	009	016	020	051	093	146	177	176	0+	177	176	0+	9		
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	005	007	022	053	098	145	176	10	0+	176	10		
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	008	023	053	097	144	11	0+	144	11		
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	008	024	053	096	12	0+	096	12	12		
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	008	023	052	13	0+	052	13	13		
14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	008	022	052	14	0+	022	14			
15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	007	15	0+	007	15			
16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	16	0+	002	16			
17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	17		
18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	18		
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	19		
20	0	818	668	442	358	290	189	122	078	049	039	031	019	012	007	004	003	001	0+	0+	0+	0+	0+	0+	0+	0+	0		
	1	165	272	368	377	370	328	270	212	159	137	117	083	058	039	026	021	007	002	0+	0+	001	0+	0+	0+	0+	1		
	2	016	053	146	189	225	271	285	274	229	217	173	137	105	078	067	028	010	003	001	0+	0+	001	0+	0+	0+	2		
	3	001	006	036	060	086	141	190	224	241	243	241	228	205	173	143	134	072	032	012	004	001	0+	001	0+	0+	3		
	4	0+	001	006	013	023	052	090	130	167	182	195	213	218	219	202	149	091	047	020	007	0+	007	020	007	0+	4		
	5	0+	0+	001	002	005	015	032	057	087	103	119	149	175	192	201	202	179	127	075	036	015	0+	036	015	0+	5		
	6	0+	0+	0+	0+	001	003	009	019	035	045	057	082	109	136	159	169	192	174	124	075	037	0+	075	037	0+	6		
	7	0+	0+	0+	0+	0+	001	002	005	012	016	022	036	055	076	100	112	164	184	166	122	074	7	0+	184	166	122	074	7
	8	0+	0+	0+	0+	0+	0+	001	003	005	007	013	022	035	051	061	114	161	180	162	120	8	0+	180	162	120	8		
	9	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	007	013	022	027	065	116	160	177	160	9	0+	177	160	9		
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	008	010	031	069	117	159	176	10	0+	176	10			
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	012	034	071	119	160	11	0+	160	11			
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	004	014	035	073	120	12	0+	120	12	12		
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	015	037	074	13	0+	074	13	13			
14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	015	037	14	0+	037	14	14				
15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	015	15	0+	015	15	15				
16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	16	0+	005	16	16				
17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	17		
18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	18		
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	19		
20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	20		
21	0	810	654	424	341	273	174	109	068	042	033	026	015	010	005	003	002	001	0+	0+	0+	0+	0+	0+	0+	0+	0		
	1	172	280	371	376	366	317	255	195	144	122	103	071	048	032	021	017	005	001	0+	0+	0+	0+	0+	0+	0+	1		
	2	017	057	155	198	233	276	284	267	234	215	196	157	121	091	066	055	022	007	002	0+	0+	002	0+	0+	0+	2		
	3	001	007	041	066	094	152	200	230	242	241	236	218	192	162	132	117	058	024	009	003	001	0+	003	001	0+	3		
	4	0+	001	008	016	027	059	100	141	177	191	202	215	216	205	187	176	113	059	026	009	003	003	003	003	003	4		
	5	0+	0+	001	003	006	018	038	065	098	115	131	161	183	197	201	199	164	109	059	026	010	0+	026	010	0+	5		
	6	0+	0+	0+	0+	001	004	011	024	043	054	067	094	122	148	169	177	188	156	057	026	0+	057	026	0+	6			
	7	0+	0+	0+	0+	0+	001	003	007	015	020	027	044	065	089	114	126	172	180	149	101	055	7	0+	101	055	7		
	8	0+	0+	0+	0+	0+	001	002	004	006	009	017	029	044	063	074	129	169	147	144	097	8	0+	144	097	8			
	9	0+	0+	0+	0+	0+	0+	0+	0+	001	002	002	005	010	018	029	036	080	132	168	170	140	9	0+	170	140	9		
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	006	011	014	041	085	134	167	168	10	0+	168	10		
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	005	018	046	089	137	168	11	0+	168	11			
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	006	021	050	093	140	12	0+	140	12	12		
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	008	023	053	097	13	0+	097	13	13		
14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	009	025	055	14	0+	055	14	14			
15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	009	026	15	0+	026	15	15			
16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	010	16	0+	010	16	16			
17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	17	0+	003	17	17			
18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	18	0+	001	18	18			
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	19		
20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	20		
21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21		

BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS (Continued)

$$P(r) = {}^nC_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r	
22	0	802	641	407	324	256	160	098	060	036	028	022	013	007	004	002	002	0+	0+	0+	0+	0+	0	
	1	178	288	373	375	360	306	241	180	130	109	090	061	041	026	017	013	004	001	0+	0+	0+	1	
	2	019	062	163	207	241	279	281	258	222	201	181	141	107	078	055	046	017	005	001	0+	0+	2	
	3	001	008	045	073	103	162	208	235	241	237	230	207	178	146	116	102	047	018	006	002	0+	3	
	4	0+	001	009	018	031	067	110	152	186	199	208	216	211	196	174	161	096	047	019	006	002	4	
	5	0+	0+	001	003	007	021	044	075	109	126	143	170	190	199	197	193	149	091	046	019	006	5	
	6	0+	0+	0+	001	005	014	029	050	063	077	106	134	159	177	183	181	139	086	043	018	6		
	7	0+	0+	0+	0+	0+	001	004	009	019	025	033	053	077	102	128	139	177	171	131	081	041	7	
	8	0+	0+	0+	0+	0+	0+	001	002	006	008	012	022	036	054	075	087	142	173	164	125	076	8	
	9	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	007	014	024	037	045	095	145	170	164	119	9	
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	005	009	015	020	053	101	148	169	154	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	001	003	005	007	025	060	107	151	168	11	
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	002	010	029	066	113	154	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	012	034	071	119	13	
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	014	037	076	076	14	
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	016	041	15	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	006	018	16	16
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	007	17	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	18	18
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	19	19
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	20	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21	21
	22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22	22
23	0	794	628	391	307	241	147	089	053	031	024	018	010	006	003	002	001	0+	0+	0+	0+	0+	0	
	1	184	295	375	372	354	294	226	166	117	097	079	053	034	021	013	010	003	001	0+	0+	0+	1	
	2	020	066	172	215	248	281	277	249	209	188	166	127	093	066	046	038	013	004	001	0+	0+	2	
	3	001	009	050	079	111	171	215	237	238	232	222	195	163	131	101	088	038	014	004	001	0+	3	
	4	0+	001	010	021	035	074	120	162	194	204	211	214	204	185	160	146	082	037	014	004	001	4	
	5	0+	0+	002	004	009	025	051	084	120	137	153	179	194	198	192	185	133	076	035	013	004	5	
	6	0+	0+	0+	001	002	006	017	034	059	073	087	118	145	168	182	185	171	122	070	032	012	6	
	7	0+	0+	0+	0+	0+	001	005	011	023	031	040	063	088	115	139	150	178	160	113	064	029	7	
	8	0+	0+	0+	0+	0+	0+	001	003	008	011	015	028	044	065	088	100	153	172	151	105	058	8	
	9	0+	0+	0+	0+	0+	0+	001	002	003	005	010	018	030	046	056	109	155	168	143	097	9	9	
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	006	012	020	026	065	117	157	164	136	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	008	010	033	074	123	159	161	11	11	
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	014	040	082	130	161	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	005	018	046	090	136	13	13
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	007	022	053	097	14	14	
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	009	026	058	15	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	011	029	16	16
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	012	17	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	18	18
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	19	19
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	20	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21	21
	22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22	22
23	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	23	23	
24	0	786	616	375	292	227	135	080	047	027	020	015	009	005	003	001	001	0+	0+	0+	0+	0+	0	
	1	190	302	375	369	347	282	213	152	105	086	070	045	028	017	010	008	002	0+	0+	0+	0+	1	
	2	022	071	180	223	255	282	272	239	196	174	153	114	081	056	038	031	010	003	001	0+	0+	2	
	3	002	011	055	086	119	180	221	239	234	225	213	183	149	117	088	075	031	010	003	001	0+	3	
	4	0+	001	012	024	040	082	129	171	200	209	213	211	196	173	146	132	069	029	010	003	001	4	
	5	0+	0+	002	005	010	029	057	093	130	147	162	185	196	195	184	176	118	062	027	009	003	5	
	6	0+	0+	0+	001	002	008	020	040	067	082	098	129	155	174	184	185	160	106	056	024	008	6	
	7	0+	0+	0+	0+	002	006	014	028	037	048	073	100	126	149	159	176	147	096	050	021	7	7	
	8	0+	0+	0+	0+	0+	001	004	010	014	019	034	053	076	100	112	160	168	136	087	044	8	8	
	9	0+	0+	0+	0+	0+	0+	001	003	004	007	013	024	038	056	067	122	161	161	126	078	9	9	
	10	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	009	016	027	033	079	130	161	155	117	10	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	006	011	014	043	089	137	161	149	11	11	
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	005	020	052	099	143	161	12	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	008	026	061	108	149	13	13
14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	003	011	032	069	117	14	14		

BINOMIAL DISTRIBUTION—INDIVIDUAL TERMS (Continued)

$$P(r) = {}_n C_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r
24	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	014	038	078	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	017	044	16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	007	021	17	
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	008	18	
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	19	
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21
	22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22
	23	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	23
	24	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	24
25	0	778	603	360	277	213	124	072	041	023	017	013	007	004	002	001	001	0+	0+	0+	0+	0+	0
	1	196	308	375	365	340	270	199	140	094	076	061	038	024	014	008	006	001	0+	0+	0+	0+	1
	2	024	075	188	231	260	282	266	228	183	161	137	101	071	048	031	025	007	002	0+	0+	0+	2
	3	002	012	060	093	127	188	226	239	229	217	203	170	136	104	076	064	024	008	002	0+	0+	3
	4	0+	001	014	027	045	090	138	179	205	211	213	206	187	161	132	118	057	022	007	002	0+	4
	5	0+	0+	002	006	012	033	065	103	140	156	170	190	196	190	175	165	103	051	020	006	002	5
	6	0+	0+	0+	001	003	010	024	047	076	092	108	139	163	179	184	183	147	091	044	017	005	6
	7	0+	0+	0+	0+	0	002	007	017	034	044	056	083	111	137	158	165	171	133	080	038	014	7
	8	0+	0+	0+	0+	0	002	005	012	017	024	041	062	087	112	124	165	161	120	070	032	8	
	9	0+	0+	0+	0+	0+	0+	001	004	006	009	017	029	046	067	078	134	163	151	108	061	9	
	10	0+	0+	0+	0+	0+	0+	0+	001	002	003	006	012	021	034	042	092	141	161	142	097	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	008	015	019	054	103	147	158	11	
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	005	007	027	065	114	151	155	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	002	011	035	076	124	13	
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	016	043	087	153	14
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	006	021	052	097	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	009	027	061	16
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	012	032	17
18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	014	18		
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	19		
20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	20	
21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21	
22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22	
23	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	23	
24	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	24	
25	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	25	

G. BINOMIAL DISTRIBUTION— CUMULATIVE TERMS

THE TABLE presents the binomial probability for r or more successes in n trials for selected values of p , the probability of a success on any one trial.

Examples and details in the use of this table for p greater than .50 are given on pages 171 and 172.

The symbol $0+$ indicates a value, positive but less than .0005.

The symbol $1-$ indicates a value, less than 1 but greater than .9995.

BINOMIAL DISTRIBUTION—CUMULATIVE TERMS

$$\text{Probability of } r \text{ or more successes in } n \text{ trials} = \sum_{r=0}^n C_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.020	.040	.078	.098	.116	.154	.190	.226	.260	.278	.294	.328	.360	.392	.422	.438	.510	.578	.640	.698	.750	1
	2	0+	0+	.002	.002	.004	.006	.010	.014	.020	.022	.026	.032	.040	.048	.058	.062	.090	.122	.160	.202	.250	2
3	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.030	.059	.115	.143	.169	.221	.271	.319	.364	.386	.407	.449	.488	.525	.561	.578	.657	.725	.784	.834	.875	1
	2	0+	.001	.005	.007	.010	.018	.028	.040	.053	.061	.069	.086	.104	.124	.145	.156	.216	.282	.352	.425	.500	2
3	3	0+	0+	0+	0+	0+	.001	.001	.002	.003	.003	.004	.006	.008	.011	.014	.016	.027	.043	.064	.091	.125	3
	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.039	.078	.151	.185	.219	.284	.344	.400	.453	.478	.502	.548	.590	.630	.666	.684	.760	.821	.870	.908	.938	1
4	2	.001	.002	.009	.014	.020	.034	.052	.073	.097	.123	.151	.181	.212	.245	.282	.318	.437	.525	.609	.688	.750	2
	3	0+	0+	0+	0+	.001	.002	.004	.006	.010	.012	.014	.020	.027	.036	.045	.051	.084	.126	.179	.241	.312	3
	4	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.001	.002	.002	.003	.004	.008	.015	.026	.041	.062	4
5	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.049	.096	.185	.226	.266	.341	.410	.472	.530	.556	.582	.629	.672	.711	.746	.763	.832	.884	.922	.950	.969	1
	2	.001	.004	.015	.023	.032	.054	.081	.112	.147	.183	.222	.263	.304	.346	.387	.422	.572	.663	.744	.812	.875	2
5	3	0+	0+	.001	.001	.002	.005	.009	.014	.022	.027	.032	.044	.058	.074	.093	.104	.163	.235	.317	.407	.500	3
	4	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.003	.004	.007	.010	.013	.016	.031	.054	.087	4
	5	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.001	.002	.005	.010	.018	.031	5
6	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.059	.114	.217	.265	.310	.394	.469	.536	.595	.623	.649	.696	.738	.775	.807	.822	.882	.925	.953	.972	.984	1
	2	.001	.006	.022	.033	.046	.077	.114	.156	.200	.224	.247	.296	.345	.394	.442	.466	.580	.681	.767	.836	.891	2
6	3	0+	0+	.001	.002	.004	.009	.016	.026	.039	.047	.056	.075	.099	.125	.154	.169	.256	.353	.456	.558	.656	3
	4	0+	0+	0+	0+	0+	.001	.001	.003	.005	.006	.007	.012	.017	.024	.033	.038	.070	.117	.179	.255	.344	4
	5	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.003	.004	.005	.011	.022	.041	.069	.109	5
6	6	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.002	.004	.008	.016	6
	7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	0	.068	.132	.249	.302	.352	.442	.522	.591	.652	.679	.705	.751	.790	.824	.854	.867	.918	.951	.972	.985	.992	1
7	1	.002	.008	.029	.044	.062	.103	.150	.201	.256	.283	.311	.368	.423	.478	.530	.555	.671	.766	.841	.898	.938	2
	2	0+	0+	.002	.004	.006	.014	.026	.042	.062	.074	.087	.115	.148	.184	.223	.244	.353	.468	.580	.684	.773	3
	3	0+	0+	0+	0+	0+	.001	.003	.005	.009	.012	.015	.023	.033	.046	.062	.071	.126	.200	.290	.392	.500	4
7	4	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.003	.005	.007	.011	.013	.029	.056	.096	.153	.227	5
	5	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.004	.009	.036	6
	6	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.002	.004	.008	.016	7
8	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.077	.149	.279	.337	.390	.487	.570	.640	.701	.728	.752	.802	.832	.863	.889	.900	.942	.968	.983	.992	.996	1
	2	.003	.010	.038	.057	.079	.130	.187	.248	.311	.343	.374	.437	.497	.554	.608	.633	.745	.831	.894	.937	.965	2
8	3	0+	0+	.003	.006	.010	.021	.038	.061	.089	.105	.123	.161	.203	.249	.297	.321	.448	.572	.685	.780	.855	3
	4	0+	0+	0+	0+	.001	.002	.005	.010	.017	.021	.027	.040	.056	.076	.100	.114	.194	.294	.406	.523	.637	4
	5	0+	0+	0+	0+	0+	0+	0+	0+	.001	.002	.003	.004	.007	.010	.016	.023	.027	.058	.106	.174	.260	5
8	6	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.011	.025	.050	.088	.145	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.004	.009	.018	.035	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.004	.008	8
9	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.086	.166	.307	.370	.427	.528	.613	.684	.743	.768	.792	.832	.866	.893	.915	.925	.960	.979	.990	.995	.998	1
	2	.003	.013	.048	.071	.098	.158	.225	.295	.366	.401	.435	.501	.564	.622	.675	.700	.804	.879	.929	.961	.980	2
9	3	0+	.001	.004	.008	.014	.030	.053	.083	.120	.141	.163	.210	.262	.316	.371	.399	.537	.663	.768	.850	.910	3
	4	0+	0+	0+	.001	.001	.004	.008	.016	.027	.034	.042	.062	.086	.114	.148	.166	.270	.391	.517	.639	.746	4
	5	0+	0+	0+	0+	0+	0+	0+	0+	.001	.002	.004	.006	.007	.012	.020	.029	.042	.049	.099	.172	.267	5
9	6	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.003	.005	.008	.010	.025	.054	.099	.166	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.004	.009	.020	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.004	.009	.020	.040	8
9	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.004	.008	9
	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	.096	.183	.335	.401	.461	.566	.651	.721	.779	.803	.825	.863	.893	.917	.936	.944	.972	.987	.994	.997	.999	1
10	2	.004	.016	.058	.086	.118	.188	.264	.342	.418	.456	.492	.561	.624	.682	.733	.756	.851	.914	.954	.977	.989	2
	3	0+	.001	.006	.012	.019	.040	.070	.109	.155	.180	.206	.263	.322	.383	.444	.474	.617	.738	.833	.900	.945	3
	4	0+	0+	0+	.001	.002	.006	.013	.024	.040	.050	.061	.088	.121	.159	.201	.224	.350	.486	.618	.734	.828	4
10	5	0+	0+	0+	0+	0+	.001	.002	.004	.007	.010	.013	.021	.033	.048	.067	.078	.150	.249	.367	.496	.623	5
	6	0+	0+	0+	0+	0+	0+	0+	0+	.001	.001	.002	.004	.006	.010	.016	.020	.047	.095	.166	.262	.377	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.002	.003	.004	.011	.026	.055	.102	7
10	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.002	.005	.012	.027	.055	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	.001	.002	.005	.011	.020	9

BINOMIAL DISTRIBUTION—CUMULATIVE TERMS (Continued)

Probability of r or more successes in n trials = $\sum_{r} {}^n C_r p^r q^{n-r}$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r
10	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	10
11	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	105	199	362	431	494	600	686	755	810	833	853	887	914	935	951	958	980	991	996	999	1-	1
	2	005	020	069	102	138	218	303	387	469	508	545	615	678	733	781	803	887	939	970	986	994	2
	3	0+	001	008	015	025	052	090	137	191	221	252	316	383	449	513	545	687	800	881	935	967	3
	4	0+	0+	001	002	003	009	019	034	056	069	085	120	161	208	260	287	430	574	704	809	887	4
	5	0+	0+	0+	0+	0+	001	003	006	012	016	021	033	050	072	099	115	210	332	467	603	726	5
	6	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	004	007	012	019	028	034	078	149	247	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	006	008	022	050	099	174	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	012	029	061	113	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006	015	033	073	124	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	006	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	11
12	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	114	215	387	460	524	632	718	784	836	858	877	908	931	949	963	968	986	994	998	999	1-	1
	2	006	023	081	118	160	249	341	431	517	557	595	664	725	778	822	842	915	958	980	992	997	2
	3	0+	002	011	020	032	065	111	167	230	264	299	370	442	511	578	609	747	849	917	958	981	3
	4	0+	0+	001	002	004	012	026	046	075	092	111	155	205	261	320	351	507	653	775	866	927	4
	5	0+	0+	0+	0+	0+	002	004	009	018	024	031	049	073	102	138	158	276	417	562	696	806	5
	6	0+	0+	0+	0+	0+	0+	001	001	003	005	006	012	019	030	045	054	118	213	335	473	613	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	007	011	014	039	085	158	261	387	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	003	009	026	057	112	194	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006	015	036	073	124	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	008	019	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	12
13	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	122	231	412	487	553	662	746	810	859	879	896	924	945	960	972	976	990	996	999	1-	1-	1
	2	007	027	093	135	181	279	379	474	561	602	640	708	766	815	856	873	936	970	987	995	998	2
	3	0+	002	014	025	039	080	134	198	270	308	346	423	498	570	636	667	798	887	942	973	989	3
	4	0+	0+	001	003	006	016	034	061	097	118	141	194	253	316	382	416	579	722	831	907	954	4
	5	0+	0+	0+	0+	001	002	006	014	026	034	044	068	099	137	182	206	346	499	647	772	867	5
	6	0+	0+	0+	0+	0+	001	002	005	008	010	018	030	046	069	080	105	284	426	573	709	799	6
	7	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	007	012	019	024	129	229	356	500	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	006	018	046	098	179	291	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	004	006	018	046	098	179	291	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	008	020	046	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	011	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	12
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	13
14	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	131	246	435	512	579	689	771	833	879	897	913	938	956	969	979	982	993	998	999	1-	1-	1
	2	008	031	106	153	204	310	415	514	603	643	681	747	802	847	884	899	953	979	992	997	999	2
	3	0+	002	017	030	048	096	158	232	311	352	393	474	552	624	689	719	839	916	960	983	994	3
	4	0+	0+	002	004	008	021	044	077	121	147	174	235	302	372	443	479	645	779	876	937	971	4
	5	0+	0+	0+	0+	001	004	009	020	036	047	059	091	130	176	230	258	416	577	721	833	910	5
	6	0+	0+	0+	0+	0+	001	004	008	012	016	027	044	066	095	112	219	359	514	663	788	884	6
	7	0+	0+	0+	0+	0+	0+	001	001	002	003	006	012	020	031	038	093	184	308	454	605	709	7
	8	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	005	008	010	031	075	150	259	395	8
	9	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	002	008	024	058	119	212	355	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006	018	043	090	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	011	029	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	006	12
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	13
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	14
15	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	140	261	458	537	605	714	794	853	896	913	927	949	965	976	984	987	995	998	1-	1-	1-	1
	2	010	035	119	171	226	340	451	552	642	681	718	781	833	874	906	920	965	986	993	998	1-	2
	3	0+	003	020	036	057	113	184	265	352	396	439	523	602	673	736	764	873	938	973	989	996	3
	4	0+	0+	002	005	010	027	056	096	148	177	209	278	352	427	502	539	703	827	909	958	982	4

BINOMIAL DISTRIBUTION—CUMULATIVE TERMS (Continued)

$$\text{Probability of } r \text{ or more successes in } n \text{ trials} = \sum_r^n C_r p^r q^{n-r}$$

[illegible]

BINOMIAL DISTRIBUTION—CUMULATIVE TERMS (Continued)

$$\text{Probability of } r \text{ or more successes in } n \text{ trials} = \sum_r^n C_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	P.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r	
19	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
	1	174	319	540	623	691	795	865	912	943	954	964	977	986	991	995	996	999	1	1	1	1	1	
	2	015	055	175	245	317	456	580	683	767	802	832	881	917	943	962	969	990	997	999	1	1	2	
	3	001	006	038	067	102	191	295	403	509	559	606	691	763	822	869	889	954	983	995	998	1	3	
	4	001	006	006	013	024	060	115	187	271	316	362	455	545	628	703	737	867	941	977	992	998	4	
	5	0+	0+	001	002	004	015	035	069	116	144	176	248	327	410	494	535	718	850	930	972	990	5	
	6	0+	0+	0+	0+	001	003	009	020	040	054	070	111	163	225	295	332	526	703	837	922	968	6	
	7	0+	0+	0+	0+	0+	002	005	020	011	016	023	041	068	103	149	175	334	519	692	827	916	7	
	8	0+	0+	0+	0+	0+	0+	0+	001	003	004	006	013	023	040	063	077	182	334	512	683	820	8	
	9	0+	0+	0+	0+	0+	0+	0+	0+	001	001	001	001	003	007	013	022	029	084	185	333	506	676	9
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	007	009	033	087	186	329	500	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	002	011	035	088	184	324	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	011	035	087	180	12
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	012	034	084	13
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	011	032		14
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	010		15
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002		16
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		17
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		18
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		19	
20	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
	1	182	332	558	642	710	811	878	922	951	961	969	981	988	993	996	997	999	1	1	1	1	1	
	2	017	060	190	264	340	483	608	711	792	824	853	898	931	954	970	976	992	998	999	1	1	2	
	3	001	007	044	075	115	212	323	433	545	595	642	725	794	849	891	909	965	988	996	999	1	3	
	4	001	001	007	016	029	071	133	213	304	352	401	497	589	671	743	775	893	956	984	995	999	4	
	5	0+	0+	001	003	006	018	043	083	137	170	206	235	370	458	544	585	762	882	949	981	994	5	
	6	0+	0+	0+	0+	001	011	026	051	067	087	136	166	266	343	383	584	755	874	945	979		6	
	7	0+	0+	0+	0+	0+	001	002	007	015	022	030	054	087	130	184	214	322	583	750	870	942	7	
	8	0+	0+	0+	0+	0+	0+	001	004	006	009	018	032	054	083	120	182	228	399	584	748	868	8	
	9	0+	0+	0+	0+	0+	0+	0+	001	001	002	005	010	019	032	041	113	238	404	586	748		9	
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	005	010	014	048	122	245	409	588		10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	004	017	053	248	249	412		11	
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	005	020	057	131	252	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	006	021	058	132	13	
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006	021	058		14	
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006	021		15	
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006		16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001		17	
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		18	
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		19		
20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		20		
21	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
	1	190	346	576	659	727	826	891	932	958	967	974	985	991	995	997	998	999	1	1	1	1	1	
	2	019	065	204	283	362	509	635	736	814	845	872	913	943	962	976	981	994	999	1	1	2		
	3	001	008	050	085	128	234	352	470	580	676	756	821	872	910	925	973	991	998	999	1	3		
	4	001	001	009	019	034	082	152	240	338	389	440	538	630	710	779	808	914	967	987	997		4	
	5	0+	0+	001	003	007	023	052	098	161	197	237	323	414	505	592	633	802	908	963	997	996	5	
	6	0+	0+	0+	0+	001	005	014	033	063	106	163	231	308	393	437	639	737	904	961	987		6	
	7	0+	0+	0+	0+	0+	001	003	009	020	029	039	068	109	160	222	256	449	643	800	904	961	7	
	8	0+	0+	0+	0+	0+	001	002	005	008	012	024	043	070	102	138	173	347	464	600	803	905	8	
	9	0+	0+	0+	0+	0+	0+	0+	001	002	003	007	014	026	044	056	148	294	476	659	808		9	
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	008	016	021	068	162	309	488	669	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	005	006	026	077	174	321	500	11		
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	009	031	085	184	332	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	011	035	091	192	13		
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	012	038	095		14	
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	013	039		15	
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	013		16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004		17	
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001		18	
19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		19		
20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		20		
21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+		21		

BINOMIAL DISTRIBUTION—CUMULATIVE TERMS (*Continued*)

$$\text{Probability of } r \text{ or more successes in } n \text{ trials} = \sum_{r=0}^n C_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r		
22	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0		
	1	198	359	593	676	744	840	902	940	964	972	978	987	993	996	998	998	1-	1-	1-	1-	1-	1		
	2	020	071	219	302	384	535	661	760	834	863	888	926	952	970	981	985	996	999	999	1-	1-	2		
	3	001	009	056	095	142	256	380	502	612	662	707	785	846	892	926	939	979	994	998	1-	1-	3		
	4	0+	001	011	022	040	094	172	267	372	425	477	578	668	746	810	838	932	975	992	998	1-	4		
	5	0+	0+	002	004	009	027	062	115	186	226	270	362	457	550	637	677	835	928	973	992	998	5		
	6	0+	0+	0+	001	002	006	018	041	077	100	127	191	267	351	439	483	687	837	928	973	992	6		
	7	0+	0+	0+	0+	0+	001	004	012	026	037	050	085	133	193	263	301	506	698	842	929	974	7		
	8	0+	0+	0+	0+	0+	001	003	008	011	017	032	056	090	135	162	329	526	710	848	933	974	8		
	9	0+	0+	0+	0+	0+	0+	001	002	003	005	010	020	036	060	075	186	353	546	724	857	974	9		
	10	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	006	012	022	030	092	208	376	565	738	10		
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	007	010	039	107	228	396	584	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	014	047	121	246	416	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	018	055	133	262	416	13	
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	006	021	062	143	262	416	14	
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	007	024	067	15	416	15	
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	008	026	16	416	16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	008	17	416	17	
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	18	416	18	
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	19	416	19	
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	20	416	20	
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21	416	21	
22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22	416	22		
23	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0			
	1	206	372	609	693	759	853	911	947	969	976	982	990	994	997	998	999	1-	1-	1-	1-	1-	1		
	2	022	077	234	321	405	559	685	781	852	880	902	937	960	975	985	988	997	999	1-	1-	1			
	3	002	011	062	105	157	278	408	533	643	692	736	810	867	909	939	951	984	996	999	1-	1			
	4	0+	001	012	026	046	107	193	295	405	460	514	615	703	778	838	863	946	982	995	999	1-			
	5	0+	0+	002	005	011	033	073	133	212	256	303	401	499	593	678	717	864	945	981	995	999	5		
	6	0+	0+	0+	001	002	008	023	050	092	119	150	222	305	395	487	532	731	869	946	981	995	6		
	7	0+	0+	0+	0+	0+	002	006	015	033	046	062	104	160	227	305	346	560	747	876	949	983	7		
	8	0+	0+	0+	0+	0+	001	004	010	015	022	042	072	113	166	196	382	586	763	885	953	983	8		
	9	0+	0+	0+	0+	0+	0+	001	003	004	007	014	027	048	078	096	229	444	612	780	895	983	9		
	10	0+	0+	0+	0+	0+	0+	0+	001	001	002	004	009	017	031	041	120	259	444	636	798	983	10		
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	005	011	015	055	142	287	472	661	983	11		
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	003	005	021	068	164	313	500	983	12		
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	007	028	081	184	339	983	13		
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	010	035	094	202	399	983	14		
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	013	041	105	15	983	15	
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	015	047	16	983	16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	017	17	983	17	
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	18	983	18	
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	19	983	19	
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	20	983	20	
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21	983	21	
22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22	983	22		
23	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	23	983	23		
24	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0			
	1	214	384	625	708	773	865	920	953	973	980	985	991	995	997	999	999	999	1-	1-	1-	1-	1		
	2	024	083	249	339	427	583	708	801	869	894	915	946	967	980	988	991	998	1-	1-	1-	1-	1		
	3	002	012	069	116	172	301	436	563	673	720	763	833	885	924	950	960	988	997	999	1-	1			
	4	0+	001	014	030	053	121	214	324	439	495	550	650	736	807	862	885	958	987	999	1-	1			
	5	0+	0+	002	006	013	039	085	153	239	287	337	439	540	634	717	753	889	958	987	996	999	5		
	6	0+	0+	0+	001	002	010	028	060	109	139	174	254	344	439	533	578	771	896	960	987	997	6		
	7	0+	0+	0+	0+	0+	002	007	019	041	057	076	126	189	264	349	393	611	789	904	964	989	7		
	8	0+	0+	0+	0+	0+	0+	002	005	013	020	028	053	089	138	199	234	435	642	808	914	968	8		
	9	0+	0+	0+	0+	0+	0+	001	004	006	009	019	036	062	099	121	275	474	672	827	924	989	9		
	10	0+	0+	0+	0+	0+	0+	0+	001	002	002	006	013	024	042	055	153	313	511	701	846	989	10		
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	008	016	021	074	183	350	546	729	989	11		
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	005	007	031	094	213	385	581	989	12		
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	002	012	042	114	242	419	989	13		
14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	016	053	134	271	419	989	14		

BINOMIAL DISTRIBUTION—CUMULATIVE TERMS (Continued)

$$\text{Probability of } r \text{ or more successes in } n \text{ trials} = \sum_r^n C_r p^r q^{n-r}$$

n	r	.01	.02	.04	.05	.06	.08	.10	.12	.14	.15	.16	.18	.20	.22	.24	.25	.30	.35	.40	.45	.50	r	
24	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	005	022	065	154	15	
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	009	027	076	16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	010	032	17	
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	011	18	
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	19	
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21
	22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22
	23	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	23
	24	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	24
25	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
	1	222	397	640	723	787	876	928	959	977	983	987	993	996	998	999	999	1-	1-	1-	1-	1-	1-	1
	2	026	089	264	358	447	605	729	820	833	907	926	955	973	984	991	993	998	1-	1-	1-	1-	2	
	3	002	013	076	127	187	323	463	591	700	746	787	853	902	936	959	968	991	998	1-	1-	1-	3	
	4	0+	001	017	034	060	135	236	352	471	529	584	683	766	832	883	904	967	990	998	1-	1-	4	
	5	0+	0+	003	007	015	045	098	173	267	318	371	477	579	672	752	786	910	968	991	998	1-	5	
	6	0+	0+	0+	001	003	012	033	071	127	162	200	288	383	482	577	622	807	917	971	991	998	6	
	7	0+	0+	0+	0+	001	003	009	024	051	070	092	149	220	301	393	439	659	827	926	974	993	7	
	8	0+	0+	0+	0+	0+	001	002	007	017	025	036	066	109	166	235	273	488	694	846	936	978	8	
	9	0+	0+	0+	0+	0+	0+	0+	002	005	008	012	025	047	079	123	149	323	533	726	866	946	9	
	10	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	008	017	033	056	071	189	370	575	758	885	10	
	11	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	006	012	022	030	098	229	414	616	788	11	
	12	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	004	009	011	044	125	268	457	655	12	
	13	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	002	003	017	060	154	306	500	13	
	14	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	001	006	025	073	193	345	14	
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	009	034	096	212	15	
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	003	013	044	115	16	
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	017	054	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	006	022	18	
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	007	19	
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	21
	22	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	22
	23	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	23
	24	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	24
25	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	0+	25	

H. POISSON DISTRIBUTION— INDIVIDUAL TERMS

THE TABLE presents individual Poisson probabilities for the number of occurrences X per unit of measurement, for selected values of m , the mean number of occurrences per unit of measurement.

A blank space is left for values less than .0005.

$$f(x) = \frac{e^{-m} m^x}{x!}$$

x	.001	.002	.003	.004	.005	.006	.007	.008	.009	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.15	x
0	999	998	997	996	995	994	993	992	991	990	980	970	961	951	942	932	923	914	905	861	0
1	001	002	003	004	005	006	007	008	009	010	020	030	038	048	057	065	074	082	090	129	1
2													001	001	002	002	003	004	005	010	2

x	.20	.25	.30	.40	.50	.60	.70	.80	.90	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	x
0	819	779	741	670	607	549	497	449	407	368	333	301	273	247	223	202	183	165	150	135	0
1	164	195	222	268	303	329	348	359	366	368	366	361	354	345	335	323	311	298	284	271	1
2	016	024	033	054	076	099	122	144	165	184	201	217	230	242	251	258	264	268	270	271	2
3	001	002	003	007	013	020	028	038	049	061	074	087	100	113	126	138	150	161	171	180	3
4				001	002	003	005	008	011	015	020	026	032	039	047	055	063	072	081	090	4
5							001	001	002	003	004	006	008	011	014	018	022	026	031	036	5
6										001	001	001	002	003	004	005	006	008	010	012	6
7														001	001	001	001	002	003	003	7
8																			001	001	8

x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	x
0	122	111	100	091	082	074	067	061	055	050	045	041	037	033	030	027	025	022	020	018	0
1	257	244	231	218	205	193	181	170	160	149	140	130	122	113	106	098	091	085	079	073	1
2	270	268	265	261	257	251	245	238	231	224	216	209	201	193	185	177	169	162	154	147	2
3	189	197	203	209	214	218	220	222	224	224	224	223	221	219	216	212	209	205	200	195	3
4	099	108	117	125	134	141	149	156	162	168	173	178	182	186	189	191	193	194	195	195	4
5	042	048	054	060	067	074	080	087	094	101	107	114	120	126	132	138	143	148	152	156	5
6	015	017	021	024	028	032	036	041	045	050	056	061	066	072	077	083	088	094	099	104	6
7	004	005	007	008	010	012	014	016	019	022	025	028	031	035	039	042	047	051	055	060	7
8	001	002	002	002	003	004	005	006	007	008	010	011	013	015	017	019	022	024	027	030	8
9				001	001	001	001	002	002	003	003	004	005	006	007	008	009	010	012	013	9
10									001	001	001	001	002	002	002	003	003	004	005	005	10
11														001	001	001	001	001	002	002	11
12																			001	001	12

POISSON DISTRIBUTION—INDIVIDUAL TERMS (Continued)

$$f(x) = \frac{e^{-m} m^x}{x!}$$

x	m																				x
	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0	
0	017	015	014	012	011	010	009	008	007	007	006	006	005	005	004	004	003	003	003	002	0
1	068	063	058	054	050	046	043	040	036	034	031	029	026	024	022	021	019	018	016	015	1
2	139	132	125	119	112	106	100	095	089	084	079	075	070	066	062	058	054	051	048	045	2
3	190	185	180	174	169	163	157	152	146	140	135	129	124	119	113	108	103	098	094	089	3
4	195	194	193	192	190	188	185	182	179	175	172	168	164	160	156	152	147	143	138	134	4
5	160	163	166	169	171	173	174	175	175	175	175	174	173	171	170	168	168	166	163	161	5
6	109	114	119	124	128	132	136	140	143	146	149	151	154	156	157	158	159	160	160	161	6
7	064	069	073	078	082	087	091	096	100	104	109	113	116	120	123	127	130	133	135	138	7
8	033	036	039	043	046	050	054	058	061	065	069	073	077	081	085	089	092	096	100	103	8
9	015	017	019	021	023	026	028	031	033	036	039	042	045	049	052	055	059	062	065	069	9
10	006	007	008	009	010	012	013	015	016	018	020	022	024	026	029	031	033	036	039	041	10
11	002	003	003	004	004	005	006	006	007	008	009	010	012	013	014	016	017	019	021	023	11
12	001	001	001	001	002	002	002	003	003	003	004	005	005	006	007	007	008	009	010	011	12
13					001	001	001	001	001	001	002	002	002	002	003	003	003	004	004	005	13
14											001	001	001	001	001	001	001	002	002	002	14
15																	001	001	001	001	15
x	m																				x
	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3	7.4	7.5	8.0	8.5	9.0	9.5	10.0	
0	002	002	002	002	002	001	001	001	001	001	001	001	001	001	001						0
1	014	013	012	011	010	009	008	008	007	006	006	005	005	005	004	003	002	001	001		1
2	042	039	036	034	032	030	028	026	024	022	021	019	018	017	016	011	007	005	003	002	2
3	085	081	077	073	069	065	062	058	055	052	049	046	044	041	039	029	021	015	011	008	3
4	129	125	121	116	112	108	103	099	095	091	087	084	080	076	073	057	044	034	025	019	4
5	158	155	152	149	145	142	138	135	131	128	124	120	117	113	109	092	075	061	048	038	5
6	160	160	159	159	157	155	153	153	151	149	147	144	142	139	137	122	107	091	076	063	6
7	140	142	144	145	146	147	148	149	149	149	149	148	147	146	140	129	117	104	090	7	
8	107	110	113	116	119	121	124	126	128	130	132	134	135	136	137	140	138	132	123	113	8
9	072	076	079	082	086	089	092	095	098	101	104	107	110	112	114	124	130	132	130	125	9
10	044	047	050	053	056	059	062	065	068	071	074	077	080	083	086	099	110	119	124	125	10
11	024	026	029	031	033	035	038	040	043	045	048	050	053	056	059	072	085	097	107	114	11
12	012	014	015	016	018	019	021	023	025	026	028	030	032	034	037	048	060	073	084	095	12
13	006	007	007	008	009	010	011	012	013	014	015	017	018	020	021	030	040	050	062	073	13
14	003	003	003	004	004	005	005	006	006	007	008	009	009	010	011	017	024	032	042	052	14
15	001	001							003	003	004	004	005	005	006	009	014	019	027	035	15
16			001	001	001	001	001	001	001	001	002	002	002	002	003	005	007	011	016	022	16
17									001	001	001	001	001	001	001	002	002	006	009	013	17
18																001	002	003	005	007	18
19																	001	001	002	004	19
20																		001	001	002	20
21																			001	001	21

I. POISSON DISTRIBUTION— CUMULATIVE TERMS

THE TABLE presents the Poisson probabilities of X or more occurrences per unit of measurement, for selected values of m , the mean number of occurrences per unit of measurement.

The symbol 1— indicates a value less than 1 but greater than .9995. A blank space is left for values less than .0005.

$$\sum_{x=0}^{\infty} \frac{e^{-m} m^x}{x!}$$

x	.001	.002	.003	.004	.005	.006	.007	.008	.009	.01	^m 1.02	.03	.04	.05	.06	.07	.08	.09	.10	.15	x
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	.001	.002	.003	.004	.005	.006	.007	.008	.009	.010	.020	.030	.039	.049	.058	.068	.077	.086	.095	.139	1
2													.001	.001	.002	.002	.003	.004	.005	.010	2
3																				.001	3

x	.20	.25	.30	.40	.50	.60	.70	.80	.90	^m 1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	x
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	.181	.221	.259	.330	.393	.451	.503	.551	.593	.632	.667	.699	.727	.753	.777	.798	.817	.835	.850	.865	1
2	.018	.026	.037	.062	.090	.122	.156	.191	.228	.264	.301	.337	.373	.408	.442	.475	.507	.537	.566	.594	2
3	.001	.002	.004	.008	.014	.023	.034	.047	.063	.080	.100	.121	.143	.167	.191	.217	.243	.269	.296	.323	3
4				.001	.002	.003	.006	.009	.013	.019	.026	.034	.043	.054	.066	.079	.093	.109	.125	.143	4
5							.001	.001	.002	.004	.005	.008	.011	.014	.019	.024	.030	.036	.044	.053	5
6										.001	.001	.002	.002	.003	.004	.006	.008	.010	.013	.017	6
7														.001	.001	.001	.002	.003	.003	.005	7
8																		.001	.001	.001	8

x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	^m 3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	x
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	.878	.889	.900	.909	.918	.926	.933	.939	.945	.950	.955	.959	.963	.967	.970	.973	.975	.978	.980	.982	1
2	.620	.645	.669	.692	.713	.733	.751	.769	.785	.801	.815	.829	.841	.853	.864	.874	.884	.893	.901	.908	2
3	.350	.377	.404	.430	.456	.482	.506	.531	.554	.577	.599	.620	.641	.660	.679	.697	.715	.731	.747	.762	3
4	.161	.181	.201	.221	.242	.264	.286	.308	.330	.353	.375	.397	.420	.442	.463	.485	.506	.527	.547	.567	4
5	.062	.072	.084	.096	.109	.123	.137	.152	.168	.185	.202	.219	.237	.256	.275	.294	.313	.332	.352	.371	5
6	.020	.025	.030	.036	.042	.049	.057	.065	.074	.084	.094	.105	.117	.129	.142	.156	.170	.184	.199	.215	6
7	.006	.007	.009	.012	.014	.017	.021	.024	.029	.034	.039	.045	.051	.058	.065	.073	.082	.091	.101	.111	7
8	.001	.002	.003	.003	.004	.005	.007	.008	.010	.012	.014	.017	.020	.023	.027	.031	.035	.040	.045	.051	8
9			.001	.001	.001	.001	.002	.002	.003	.004	.005	.006	.007	.008	.010	.012	.014	.016	.019	.021	9
10							.001	.001	.001	.001	.001	.002	.002	.003	.003	.004	.005	.006	.007	.008	10
11													.001	.001	.001	.001	.002	.002	.002	.003	11
12																		.001	.001	.001	12

POISSON DISTRIBUTION—CUMULATIVE TERMS (Continued)

$$\sum_{x=0}^{\infty} \frac{e^{-m} m^x}{x!}$$

	m																				
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0	x
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	983	985	986	988	989	990	991	992	993	993	994	994	995	995	996	997	997	997	997	998	1
2	915	922	928	934	939	944	948	952	956	960	963	966	969	971	973	976	978	979	981	983	2
3	776	790	803	815	826	837	848	857	867	875	884	891	898	905	912	918	923	928	933	938	3
4	586	605	623	641	658	674	690	706	721	735	749	762	775	787	798	809	820	830	840	849	4
5	391	410	430	449	468	487	505	524	542	560	577	594	610	627	642	658	673	687	701	715	5
6	231	247	263	280	297	314	332	349	366	384	402	419	437	454	471	488	505	522	538	554	6
7	121	133	144	156	169	182	195	209	223	238	253	268	283	298	314	330	346	362	378	394	7
8	057	064	071	079	087	095	104	113	123	133	144	155	167	178	191	203	216	229	242	256	8
9	024	028	032	036	040	045	050	056	062	068	075	082	089	097	106	114	123	133	143	153	9
10	010	011	013	015	017	020	022	025	028	032	036	040	044	049	054	059	065	071	077	084	10
11	003	004	005	006	007	008	009	010	012	014	016	018	020	023	025	028	031	035	039	042	11
12	001	001	002	002	002	003	003	004	005	005	006	007	008	010	011	012	014	016	018	020	12
13			001	001	001	001	001	001	002	002	002	003	003	004	004	005	006	007	008	009	13
14									001	001	001	001	001	001	002	002	002	003	003	004	14
15															001	001	001	001	001	001	15
16																			001	001	16

x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0	^m 7.1	7.2	7.3	7.4	7.5	8.0	8.5	9.0	9.5	10.0	x
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	998	998	998	998	998	999	999	999	999	999	999	999	999	999	999	1-	1-	1-	1-	1-	2
2	984	985	987	988	989	990	991	991	992	993	993	994	995	995	997	998	999	999	999	1-	2
3	942	946	950	954	957	960	963	966	968	970	973	975	976	978	980	986	991	994	996	997	3
4	857	866	874	881	888	895	901	907	913	918	923	928	933	937	941	958	970	979	985	990	4
5	728	741	753	765	776	787	798	808	818	827	836	844	853	860	868	900	926	945	960	971	5
6	570	586	601	616	631	645	659	673	686	699	712	724	736	747	759	809	850	884	911	933	6
7	410	426	442	458	473	489	505	520	535	550	565	580	594	608	622	687	744	793	835	870	7
8	270	284	298	313	327	342	357	372	386	401	416	431	446	461	475	547	614	676	731	780	8
9	163	174	185	197	208	220	233	245	258	271	284	297	311	324	338	407	477	544	608	667	9
10	091	098	106	114	123	131	140	150	151	170	180	190	201	212	224	283	347	413	478	542	10
11	047	051	056	061	067	073	079	085	092	099	106	113	121	129	138	184	237	294	355	417	11
12	022	025	028	031	034	037	041	045	049	053	058	063	068	074	079	112	151	197	248	303	12
13	010	011	013	014	016	018	020	022	024	027	030	033	036	039	043	064	091	124	164	208	13
14	004	005	005	006	007	008	009	010	011	013	014	016	018	020	022	034	051	074	102	136	14
15	002	002	002	003	003	003	004	004	005	006	006	007	008	009	010	017	027	041	060	083	15
16	001	001	001	001	001	001	002	002	002	002	003	003	004	004	005	008	014	022	033	049	16
17						001	001	001	001	001	001	001	001	002	002	004	007	011	018	027	17
18																002	003	005	009	014	18
19																001	001	002	004	007	19
20																	001	001	002	003	20
21																			001	002	21
22																			001	001	22

J. VALUES OF t

THE VALUE t describes the sampling distribution of a deviation from a population value divided by the standard error.

Probabilities in the heading refer to the sum of the two-tailed areas under the curve that lie outside the points $\pm t$. (For a single tail divide the probability by 2.) Degrees of freedom are listed in the first column.

Example: In the distribution of the means of samples of size $n = 10$, $df = n - 1 = 9$; then .05 of the area under the curve falls in the two tails outside the interval $t = \pm 2.262$. The last row shows the corresponding areas under the normal curve.

PROBABILITY (P)

df	.20	.10	.05	.02	.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365 ✓	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
∞	1.28155	1.64485	1.95996	2.32634	2.57582

Reprinted from Table IV, p. 174, of R. A. Fisher, *Statistical Methods for Research Workers* (11th ed.), published by Oliver and Boyd, Ltd. Edinburgh, by permission of the author and publishers.

K. SUMS OF SQUARES AND FOURTH POWERS USED IN TREND FITTING

THIS TABLE gives the values of Σx^2 and Σx^4 needed to find the constants in secular trend equations fitted by least squares, where the x origin is centered at the midpoint in time. Use the left half of the table for an odd number of years, where the x unit is one year. Use the right half of the table for an even number of years, where the x unit is six months, and the years are numbered 1, 3, 5, . . . and $-1, -3, -5$. . . from the origin. The sum includes the powers of negative as well as positive values of x . For example, $N = 51$ includes integer values of x from -25 to 25 , and $N = 50$ includes odd-numbered values of x from -49 to 49 .

FOR ODD NUMBER OF YEARS x UNIT IS 1 YEAR			FOR EVEN NUMBER OF YEARS x UNIT IS 6 MONTHS		
N	Σx^2	Σx^4	N	Σx^2	Σx^4
3	2	2	2	2	2
5	10	34	4	20	164
7	28	196	6	70	1 414
9	60	708	8	168	6 216
11	110	1 958	10	330	19 338
13	182	4 550	12	572	48 620
15	280	9 352	14	910	105 742
17	408	17 544	16	1 360	206 992
19	570	30 666	18	1 938	374 034
21	770	50 666	20	2 660	634 676
23	1 012	79 948	22	3 542	1 023 638
25	1 300	121 420	24	4 600	1 583 320
27	1 638	178 542	26	5 850	2 364 570
29	2 030	255 374	28	7 308	3 427 452
31	2 480	356 624	30	8 990	4 842 014
33	2 992	469 696	32	10 912	6 689 056
35	3 570	654 738	34	13 090	9 060 898
37	4 218	864 690	36	15 540	12 062 148
39	4 940	1 125 332	38	18 278	15 810 470
41	5 740	1 445 332	40	21 320	20 437 352
43	6 622	1 834 294	42	24 682	26 088 874
45	7 590	2 302 806	44	28 380	32 926 476
47	8 628	2 862 488	46	32 430	41 127 726
49	9 800	3 526 040	48	36 848	50 887 088
51	11 050	4 307 290	50	41 650	62 416 690
53	12 402	5 221 242	52	46 852	75 947 092
55	13 860	6 284 124	54	52 470	91 728 054
57	15 428	7 513 436	56	58 520	110 029 304
59	17 110	8 927 998	58	65 018	131 141 306
61	18 910	10 547 998	60	71 980	155 376 028

APPENDIX L. RANDOM NUMBERS

22 17 68 65 84	68 95 23 92 35	87 02 22 57 51	61 09 43 95 06	58 24 82 03 47
19 36 27 59 46	13 79 93 37 55	39 77 32 77 09	85 52 05 30 62	47 83 51 62 74
16 77 23 02 77	09 61 87 25 21	28 06 24 25 93	16 71 13 59 78	23 05 47 47 25
78 43 76 71 61	20 44 90 32 64	97 67 63 99 61	46 38 03 93 22	69 81 21 99 21
03 28 28 26 08	73 37 32 04 05	69 30 16 09 05	88 69 58 28 99	35 07 44 75 47
93 22 53 64 39	07 10 63 76 35	87 03 04 79 88	08 13 13 85 51	55 34 57 72 69
78 76 58 54 74	92 38 70 96 92	52 06 79 79 45	82 63 18 27 44	69 66 92 19 09
23 68 35 26 00	99 53 93 61 28	52 70 05 48 34	56 65 05 61 86	90 92 10 70 80
15 39 25 70 99	93 86 52 77 65	15 33 59 05 28	22 87 26 07 47	86 96 98 29 06
58 71 96 30 24	18 46 23 34 27	85 13 99 24 44	49 18 09 79 49	74 16 32 23 02
57 35 27 33 72	24 53 63 94 09	41 10 76 47 91	44 04 95 49 66	39 60 04 59 81
48 50 86 54 48	22 06 34 72 52	82 21 15 65 20	33 29 94 71 11	15 91 29 12 03
61 96 48 95 03	07 16 39 33 66	98 56 10 56 79	77 21 30 27 12	90 49 22 23 62
36 93 89 41 26	29 70 83 63 51	99 74 20 52 36	87 09 41 15 09	98 60 16 03 03
18 87 00 42 31	57 90 12 02 07	23 47 37 17 31	54 08 01 88 63	39 41 88 92 10
88 56 53 27 59	33 35 72 67 47	77 34 55 45 70	08 18 27 38 90	16 95 86 70 75
09 72 95 84 29	49 41 31 06 70	42 38 06 45 18	54 84 73 31 65	52 53 37 97 15
12 96 88 17 31	65 19 69 02 83	60 75 86 90 68	24 64 19 35 51	56 61 87 39 12
85 94 57 24 16	92 09 84 38 76	22 00 27 69 85	29 81 94 78 70	21 94 47 90 12
38 64 43 59 98	98 77 87 68 07	91 51 67 62 44	40 98 05 93 78	23 32 65 41 18
53 44 09 42 72	00 41 86 79 79	68 47 22 00 20	35 55 31 51 51	00 83 63 22 55
40 76 66 26 84	57 99 99 90 37	36 63 32 08 58	37 40 13 68 97	87 64 81 07 83
02 17 79 18 05	12 59 52 57 02	22 07 90 47 03	28 14 11 30 79	20 69 22 40 98
95 17 82 06 53	31 51 10 96 46	92 06 88 07 77	56 11 50 81 69	40 23 72 51 39
35 76 22 42 92	96 11 83 44 80	34 68 35 48 77	33 42 40 90 60	73 96 53 97 86
26 29 13 56 41	85 47 04 66 08	34 72 57 59 13	82 43 80 46 15	38 26 61 70 04
77 80 20 75 82	72 82 32 99 90	63 95 73 76 63	89 73 44 99 05	48 67 26 43 18
46 40 66 44 52	91 36 74 43 53	30 82 13 54 00	78 45 63 98 35	55 03 36 67 68
37 56 08 18 09	77 53 84 46 47	31 91 18 95 58	24 16 74 11 53	44 10 13 85 57
61 65 61 68 66	37 27 47 39 19	84 83 70 07 48	53 21 40 06 71	95 06 79 88 54
93 43 69 64 07	34 18 04 52 35	56 27 09 24 86	61 85 53 83 45	19 90 70 99 00
21 96 60 12 99	11 20 99 45 18	48 13 93 55 34	18 37 79 49 90	65 97 38 20 46
95 20 47 97 97	27 37 83 28 71	00 06 41 41 74	45 89 09 39 84	51 67 11 52 49
97 86 21 78 73	10 65 81 92 59	58 76 17 14 97	04 76 62 16 17	17 95 70 45 80
69 92 06 34 13	59 71 74 17 32	27 55 10 24 19	23 71 82 13 74	63 52 52 01 41

RANDOM NUMBERS (Continued)

04 31 17 21 56	33 73 99 19 87	26 72 39 27 67	53 77 57 68 93	60 61 97 22 61
61 06 98 03 91	87 14 77 43 96	43 00 65 98 50	45 60 33 01 07	98 99 46 50 47
85 93 85 86 88	72 87 08 62 40	16 06 10 89 20	23 21 34 74 97	76 38 03 29 63
21 74 32 47 45	73 96 07 94 52	09 65 90 77 47	25 76 16 19 33	53 05 70 53 30
15 69 53 82 80	79 96 23 53 10	65 39 07 16 29	45 33 02 43 70	02 87 40 41 45
02 89 08 04 49	20 21 14 68 86	87 63 93 95 17	11 29 01 95 80	35 14 97 35 33
87 18 15 89 79	85 43 01 72 73	08 61 74 51 69	89 74 39 82 15	94 51 33 41 67
98 83 71 94 22	59 97 50 99 52	08 52 85 08 40	87 80 61 65 31	91 51 80 32 44
10 08 58 21 66	72 68 49 29 31	89 85 84 46 06	59 73 19 85 23	65 09 29 75 63
47 90 56 10 08	88 02 84 27 83	42 29 72 23 19	66 56 45 65 79	20 71 53 20 25
22 85 61 68 90	49 64 92 85 44	16 40 12 89 88	50 14 49 81 06	01 82 77 45 12
67 80 43 79 33	12 83 11 41 16	25 58 19 68 70	77 02 54 00 52	53 43 37 15 26
27 62 50 96 72	79 44 61 40 15	14 53 40 65 39	27 31 58 50 28	11 39 03 34 25
33 78 80 87 15	38 30 06 38 21	14 47 47 07 26	54 96 87 53 32	40 36 40 96 76
13 13 92 66 99	47 24 49 57 74	32 25 43 62 17	10 97 11 69 84	99 63 22 32 98
10 27 53 96 23	71 50 54 36 23	54 31 04 82 98	04 14 12 15 09	26 78 25 47 47
28 41 50 61 88	64 85 27 20 18	83 36 36 05 56	39 71 65 09 62	94 76 62 11 89
34 21 42 57 02	59 19 18 97 48	80 30 03 30 98	05 24 67 70 07	84 97 50 87 46
61 81 77 23 23	82 82 11 54 08	53 28 70 58 96	44 07 39 55 43	42 34 43 39 28
61 15 18 13 54	16 86 20 26 88	90 74 80 55 09	14 53 90 51 17	52 01 63 01 59
91 76 21 64 64	44 91 13 32 97	75 31 62 66 54	84 80 32 75 77	56 08 25 70 29
00 97 79 08 06	37 30 28 59 85	53 56 68 53 40	01 74 39 59 73	30 19 99 85 48
36 46 18 34 94	75 20 80 27 77	78 91 69 16 00	08 43 18 73 68	67 69 61 34 25
88 98 99 60 50	65 95 79 42 94	93 62 40 89 96	43 56 47 71 66	46 76 29 67 02
04 37 59 87 21	05 02 03 24 17	47 97 81 56 51	92 34 86 01 82	55 51 33 12 91
63 62 06 34 41	94 21 78 55 09	72 76 45 16 94	29 95 81 83 83	79 88 01 97 30
78 47 23 53 90	34 41 92 45 71	09 23 70 70 07	12 38 92 79 43	14 85 11 47 23
87 68 62 15 43	53 14 36 59 25	54 47 33 70 15	59 24 48 40 35	50 03 42 99 36
47 60 92 10 77	88 59 53 11 52	66 25 69 07 04	48 68 64 71 06	61 65 70 22 12
56 88 87 59 41	65 28 04 67 53	95 79 88 37 31	50 41 06 94 76	81 83 17 16 33
02 57 45 86 67	73 43 07 34 48	44 26 87 93 29	77 09 61 67 84	06 69 44 77 75
31 54 14 13 17	48 62 11 90 60	68 12 93 64 28	46 24 79 16 76	14 60 25 51 01
28 50 16 43 36	28 97 85 58 99	67 22 52 76 23	24 70 36 54 54	59 28 61 71 96
63 29 62 66 50	02 63 45 52 38	67 63 47 54 75	83 24 78 43 20	92 63 13 47 48
45 65 58 26 51	76 96 59 38 72	86 57 45 71 46	44 67 76 14 55	44 88 01 62 12
39 65 36 63 70	77 45 85 50 51	74 13 39 35 22	30 53 36 02 95	49 34 88 73 61
73 71 98 16 04	29 18 94 51 23	76 51 94 84 86	79 93 96 38 63	08 58 25 58 94
72 20 56 20 11	72 65 71 08 86	79 57 95 13 91	97 48 72 66 48	09 71 17 24 89
75 17 26 99 76	89 37 20 70 01	77 31 61 95 46	26 97 05 73 51	53 33 18 72 87
37 48 60 82 29	81 30 15 39 14	48 38 75 93 29	06 87 37 78 48	45 56 00 84 47
68 08 02 80 72	83 71 46 30 49	89 17 95 88 29	02 39 56 03 46	97 74 06 56 17
14 23 98 61 67	70 52 85 01 50	01 84 02 78 43	10 62 98 19 41	18 83 99 47 99
49 08 96 21 44	25 27 99 41 28	07 41 08 34 66	19 42 74 39 91	41 96 53 78 72
78 37 06 08 43	63 61 62 42 29	39 68 95 10 96	09 24 23 00 62	56 12 80 73 16
37 21 34 17 68	68 96 83 23 56	32 84 60 15 31	44 73 67 34 77	91 15 79 74 58

RANDOM NUMBERS (*Continued*)

14 29 09 34 04	87 83 07 55 07	76 58 30 83 64	87 29 25 58 84	86 50 60 00 25
58 43 28 06 36	49 52 83 51 14	47 56 91 29 34	05 87 31 06 95	12 45 57 09 09
10 43 67 29 70	80 62 80 03 42	10 80 21 38 84	90 56 35 03 09	43 12 74 49 14
44 38 88 39 54	86 97 37 44 22	00 95 01 31 76	17 16 29 56 63	38 78 94 49 81
90 69 59 19 51	85 39 52 85 13	07 28 37 07 61	11 16 36 27 03	78 86 72 04 95
41 47 10 25 62	97 05 31 03 61	20 26 36 31 62	68 69 86 95 44	84 95 48 46 45
91 94 14 63 19	75 89 11 47 11	31 56 34 19 09	79 57 92 36 59	14 93 87 81 40
80 06 54 18 66	09 18 94 06 19	98 40 07 17 81	22 45 44 84 11	24 62 20 42 31
67 72 77 63 48	84 08 31 55 58	24 33 45 77 58	80 45 67 93 82	75 70 16 08 24
59 40 24 13 27	79 26 88 86 30	01 31 60 10 39	53 58 47 70 93	85 81 56 39 38
05 90 35 89 95	01 61 16 96 94	50 78 13 69 36	37 68 53 37 31	71 26 35 03 71
44 43 80 69 98	46 68 05 14 82	90 78 50 05 62	77 79 13 57 44	59 60 10 39 66
61 81 31 96 82	00 57 25 60 59	46 72 60 18 77	55 66 12 62 11	08 99 55 64 57
42 88 07 10 05	24 98 65 63 21	47 21 61 88 32	27 80 30 21 60	10 92 35 36 12
77 94 30 05 39	28 10 99 00 27	12 73 73 99 12	49 99 57 94 82	96 88 57 17 91
78 83 19 76 16	94 11 68 84 26	23 54 20 86 85	23 86 66 99 07	36 37 34 92 09
87 76 59 61 81	43 63 64 61 61	65 76 36 95 90	18 48 27 45 68	27 23 65 30 72
91 43 05 96 47	55 78 99 95 24	37 55 85 78 78	01 48 41 19 10	35 19 54 07 73
84 97 77 72 73	09 62 06 65 72	87 12 49 03 60	41 15 20 76 27	50 47 02 29 16
87 41 60 76 83	44 88 96 07 80	83 05 83 38 96	73 70 66 81 90	30 56 10 48 59

Source: This table is reprinted with permission from Random Numbers III and IV of Table XXXIII of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (Edenburg: Oliver & Boyd, Ltd.).

INDEX

INDEX

A

- Acceptance sampling, 675-78
- Analysis; *see* Statistical analysis
- Antilogarithms, 691
- Area sampling, 334, 346
- Arithmetic line charts, 49-51
- Arithmetic mean
 - attribute data, 98, 99
 - characteristics of, 108
 - defined, 94, 95
 - error of grouping, 97, 98
 - formula, 109
 - grouped data, 96-98
 - open-end distributions, 98
 - short-cut methods, 98
 - tests of differences between, 289-94
 - ungrouped data, 95, 96
 - weighted, 96
- Attributes, 67
- Audimeters, 30
- Automation, 36
- Average deviation, 120-22
- Averages
 - arithmetic mean, 94-99; *see also* Arithmetic mean
 - characteristics of, 107-9
 - characteristics of frequency distributions, 114-16
 - formulas summarized, 109
 - geometric mean, 104-6; *see also* Geometric mean
 - median, 99-102; *see also* Median
 - mode, 102-4; *see also* Mode
 - modified mean, 102
 - selection of one to use, 106, 107
 - types of, 94

B

- Bar charts, 57-60, 62
- Barlow's Tables, 123
- Barron's, 20, 430
- Bayes' theorem
 - normal distributions
 - assumptions underlying, 391
 - evaluation of sampling information, 384-87
 - examples as illustration, 381-84, 385-87
 - factors influencing expected value of sample information, 387, 391
 - formulas, 391, 392
 - optimal sample size, 387-91

Bayes' theorem—*Cont.*

- population from which sample drawn, 378, 390
- posterior distribution, determination of, 377-84, 391
- prior distribution, 378, 379, 390
- sample means, 379
- probability distributions
 - binomial sampling, 362, 363
 - classical approach distinguished, 370, 371
 - decision-making, 363-67
 - economic analysis after sampling, 365-67
 - economic analysis before sampling, 364, 365
 - expected value of sample information, 367-70, 372
 - posterior, 357, 358, 363-67, 371
 - prior, 357-62, 371
- Bent coin example, 166-68
- Bias, 9
 - ratio estimate, 332
 - sampling, 250, 251
 - weighting of index numbers, 443, 444, 456
- Bibliographic Index*, 20
- Bidding model, 397-99, 414
- Binomial distribution
 - assumptions underlying, 169, 170, 184
 - bent coin example, 166-68
 - cumulative terms, 715-21
 - defined, 183
 - examples of, 166-72
 - individual terms, 708-14
 - normal approximation to, 179-82
 - Poisson as approximation to, 175, 176
 - probability formula, 168, 169
 - proportions, 304, 305, 311
 - tables of, 170, 171
- Binomial probabilities, Bayes, theorem for revision of, 362, 363
- Brand loyalty, 149, 51
- Brookings Institution, 7
- Bulletin of the Public Affairs Information Service*, 20
- Bus Facts*, 23
- Business
 - statistical activities in, 5, 6
 - statistical approach to problems of, 4-6
- Business Cycle Developments*, 427, 542
- Business cycles; *see* Cyclical fluctuations

- Business fluctuations, types of, 463-65, 495
 cyclical; *see* Cyclical fluctuations
 irregular movements; *see* Irregular fluctuations
 seasonal variation; *see* Seasonal variation
 secular trend; *see* Secular trend
- Business forecasting
 correlation of aggregates, validity of, 646-48, 651, 652
 cyclical fluctuations, 540-45, 547
 sales, 648-50, 652
 seasonal variation, 523-27
 statistical methods in, 545, 546
- Business indicators, 427, 428
Business Periodicals Index, 20
Business Week, 20
- C**
- Calendar variation, 503-5, 525
- Charts
 advantages over tables, 42, 60
 analysis through use of, 46
 arithmetic line, 49-51
 bar, 57-60, 62
 component parts bars, 59, 60
 frequency distributions, 78-81; *see also*
 Frequency distributions
 presentation of data through, 46-48
 probability distributions, 154, 155
 proportions of, 48, 49
 purposes of, 46, 61
 ratio, 51-57, 61; *see also* Ratio charts
 scale distortion, 48, 49
 types of, 49, 61
Canadian Statistical Review, 19
Catalog, 20
Census of Manufacturers, 22
Census of Population, 27
 Census taking, 27
 Central limit theorem, 259, 271
 Cleveland Trust Company, 19
- Cluster sampling
 advantages of, 334, 335
 area sampling, 334, 346
 defined, 334, 346
 disadvantages of, 335
 elementary sampling units, 334
 formulas, 336-39, 347
 notation of, 336-39
 primary sampling units, 334
 serpentine sequence, 335, 346
 subsampling, 336
 systematic selection, 335, 336
- Coefficient of dispersion, 133
 Coefficient of multiple correlation, 600, 601, 612
 Coefficient of skewness, 134
 Colinearity, 610-12
- Collection of data
 original data, 26-36; *see also* Original statistical data
 research sources, 18-26; *see also* Research sources
- Composite index numbers
 aggregative method, 435-38, 456
 average of relatives method, 433-35, 438, 456
 construction of, 432, 455, 456
 formulas for computing, 437, 438
 weights, necessity of, 432, 433, 456
Commercial and Financial Chronicle, 20
- Conditional probability, 142-44, 157
- Confidence intervals, 250, 265-67
 coefficient, selection of, 267, 268, 271, 272
 difference between sample means, 293-95
 errors in, 268
 proportions, 306, 307, 311
 sampling and regression analysis, 567-71, 583
 small samples, 302
- Consumer price index
 basis for, 448
 defined, 447
 limitations of, 450, 451
 uses of, 449-51, 456
- Continuous distributions
 defined, 158
 expected value of, 158, 159
 variance, 159
- Control charts for variables, 660-75, 677
- Correlation, defined, 551
- Correlation between two variables; *see*
 Multiple correlation; Regression analysis; *and* Simple correlation
- Correlation coefficient, 571-74
 graphic analysis, 574-76
 sampling error of, 576, 577
- Cumulative frequency distributions, 82, 83
- Current Statistics*, 19
- Curvilinear correlation, 553, 554, 581
- Curvilinear regression, 631, 650
 graphic analysis, 631, 650
 multiple regression, 632-35
 simple regression, 631, 632
 index of correlation, 638, 650
 mathematical curves, 635
 homoscedasticity, 638
 logarithms, 636, 637, 650
 parabola, 635, 636, 650
 transformations, 637, 638, 650
 standard error of estimate, 638, 650
 statistical inference, 639
 when to make use of, 639, 640, 651
- Curvilinear relationships, regression analysis, 579, 580

Cyclical fluctuations

- arithmetic adjustment for measurement, 536-38
- average duration of run, 543, 547
- defined, 546
- diffusion indexes, 542, 543, 547
- electronic computer for measurement, 538, 539, 546
- exponentially weighted moving averages, 541, 547
- forecasting of, 540-45, 547
- graphic method of measurement, 536, 546
- importance of, 532, 546
- lead and lag indicators, 541, 542, 547
- length of, 532, 533, 546
- methods of measurement of, 535-39, 546
- reasons for measurement of, 534, 535
- surveys of anticipations data, 543-45, 547

Cyclical rhythm, 503

D

Data; *see* Statistical data*Data Processing Digest*, 36

Decision-making

- Bayes' theorem; *see* Bayes' theorem
- certainty in, 189
- economic factors only as basis, 194, 195
- electronic computer, role of, 4
- expected monetary value as criterion, 195-99
- expected value of perfect information, 226-28, 237, 238, 241
 - alternative method for determination of, 228
 - illustration, 229, 230
- formulas, 242
- how to do it, 191-99
- hypothetical example, 191 ff.
- importance of, 1
- linear profit functions, 230-33, 242
 - opportunity loss functions, 233, 234, 242
- normal distribution in, 234, 235, 242
 - opportunity loss and, 235-41
- oil-drilling example, 197-99
- opportunity loss, 224-26, 233-41
- payoff table, 194, 195
- posterior, 357, 363-67, 371
- prior, 357, 358, 371
- probabilities only as basis, 193, 194
- probability models useful in; *see* Probability models
- procedure for, 213
- profit under certainty, 228
- risk in, 211-13

Decision-making—*Cont.*

- sequential decisions, 202-11, 213; *see also* Decision trees
- states of the world, 191
- statistics in, function of, 3, 4
- subjective probabilities and, 199-202
- transportation problem, 189
- uncertainty in, 190
 - additional information, value of, 224 ff.
 - derivation of utility curves for, 214-17

- utility curves, 214-17
- utility of money, 211-14

Decision trees

- analysis using, 204-7, 214
 - more complex situation, 207-11
- sequence of decisions, 202-4, 213
- working backward on, 204-7, 210, 211, 214

Deductive reasoning, 10, 11

Definitions

- arithmetic mean, 94, 95
- attributes, 67
- binomial distribution, 183
- cluster sampling, 334, 346
- composite index, 433
- consumer price index, 447
- continuous distribution, 158
- correlation, 551
- cyclical fluctuations, 546
- dispersion, 116
- frequency distribution, 71, 72
- geometric mean, 104
- index numbers, 427, 455
- industrial production index, 453
- irregular fluctuations, 546
- judgment sampling, 343
- logarithms, 1
- matrix, 620
- mean deviation, 120
- median, 99
- mode, 102, 103
- modified mean, 102
- multiple correlation, 551, 552
- multiple regression, 611
- nonprobability sampling, 340, 341, 346
- normal distribution, 184
- probability, 140, 157
- probability distributions, 153, 158
- probability sampling, 316, 317, 345
- probability theory, 140
- proportions, 304
- quality control, 677
- quartile deviation, 118
- quota sampling, 341
- range, 117
- ratios, 68
- regression, 551

Definitions—*Cont.*

- replicated sampling, 339, 346
- seasonal variation, 507, 525
- secular trend, 465, 495
- simple correlation, 551, 581
- skewness, 133
- statistical analysis, 1, 15
- statistical data, 1
- statistical independence, 144–46, 157
- statistical inference, 2, 249, 271
- stratified sampling, 319, 345
- subjective probability, 142, 157
- tables, 42
- variables, 67
- variance, 123
- wholesale price index, 451

Dice rolling, 148, 149

Diffusion indexes, 542, 543, 547

Dispersion

- aid in description of data, 134
- average deviation, 120–22
- characteristics of frequency distributions, 114–16
- characteristics of measures of, 131, 132
- coefficient of, 133
- comparison of, 134, 135
- defined, 116
- formulas of measures summaries, 135, 136
- mean deviation, 120
 - characteristics of, 132
 - defined, 120
 - formula for, 135, 136
 - grouped data, 122
 - short method, 126–28
 - ungrouped data, 121, 122
- measure to use, selection of, 130, 131
- measures of, 116, 117
- purposes of measuring, 116, 117
- quartile deviation, 118
 - characteristics of, 131, 132
 - defined, 118
 - formula for, 135
 - grouped data, 119, 120
 - ungrouped data, 118, 119
- range, 117, 118
 - characteristics of, 131
 - formula for, 135
- relation between measures of, 129–32
- relative measures of, 132, 133
 - formula for, 136
- sampling errors, measurement of, 135
- standard deviation
 - formulas for, 135, 136
 - units of, 133
- skewness
 - coefficient of, 134
 - defined, 133
 - formula for, 136

Dispersion—*Cont.*

- standard, provision of, 135
- standard deviation
 - basis of, 122
 - characteristics of, 132
 - grouped data, 124–28
 - short methods, 124–28
 - ungrouped data, 123, 124
- uses of measures of, 134, 135

Dollar = value indexes, 437

Double-sampling plan, 676

Dun's Review and Modern Industry, 19, 118

E

Economic Almanac, 19*Economic Indicators*, 19, 427

Economics, uses of statistics in, 7

Economist, 20

- Electronic computer, 3; *see also* Electronic data processing
 - cyclical fluctuations measured by, 538, 539, 546
 - decision-making process, 4
 - functions of, 35
 - irregular fluctuations measured by, 538, 539
 - multiple regression, 603–10, 612
 - seasonal variation measurement, 518–21, 526

Electronic data processing; 35, 36; *see also* Electronic computer

Expected value of sample information

- Bayes' theorem, 367–70, 372
- factors influencing, 387, 391

F

Facts from Figures, 8*Federal Reserve Bulletin*, 19, 427*Federal Reserve Chart Book: Financial and Business Statistics*, 19

First National City Bank of New York, 19

Forecasting; *see* Business forecasting

Formulas

- averages, 109
- Bayes' theorem for normal distributions, 391, 392
- binomial probability, 168, 169
- cluster sampling, 336–39, 347
- composite indexes, computation of, 437, 438
- decision-making, 242
- dispersion measures, 135, 136
- Poisson distribution, 173
- ratio estimation, 347
- ratio sampling, 330–32
- replicated sampling, 347
- simple random sample, 347
- stratified sampling, 347
 - least-cost allocation, 347
 - optimal allocation, 347

- Frequency curves, 84
 - types of, 85-87
- Frequency distributions
 - array of variables, 72, 73, 87
 - average, 114-16
 - characteristics of, 114-16
 - charts of, 78-81
 - advantages of, 81
 - class intervals, 75-77
 - comparison of two, 81
 - cumulative, 82, 83
 - defined, 71, 72, 87
 - dispersion, 114-16; *see also* Dispersion
 - frequency polygon, 79, 80, 88
 - advantages of, 80
 - grouping data into classes, 73-75, 87
 - histogram, 78, 79, 88
 - advantages of, 80
 - homogeneity, 72
 - kurtosis, 114-16
 - ogive, 82, 83, 88
 - percentage, 77, 78
 - relative, 77, 78
 - skewness, 114-16
 - variables in, 72
- Frequency polygon, 79, 80, 88
 - advantages of, 80
- Full Employment Act of 1946, 7

G

- Generalization, 2, 9, 10
- Geometric mean
 - characteristics of, 109
 - defined, 104
 - formula, 109
 - grouped data, 105, 106
 - ungrouped data, 105
- Girshick, M. A., 4
- Glossary of symbols, 685-88
- Government statistical analysis, use of, 7
- Graphic interpolation, 102
- Graphs; *see* Charts
- Gross national product, rate of growth, 466, 467
- Guide to Industrial Statistics*, 20
- Guide to U.S. Government Statistics*, 20

H

- Histogram, 78, 79, 88
 - advantages of, 80
- Homoscedasticity, 638
- Hypergeometric distribution, 170
- Hypotheses, tests of; *see* Tests of hypotheses

I

- Index numbers
 - advantages of, 428
 - base period, choice of, 441, 442, 456
 - basic methods of constructing, 431-38

Index numbers—*Cont.*

- census years, inclusion of, 442
- changing base period of, 445, 446, 456
- comparability with others, 441, 442, 456
- composite, 432-38, 455, 456; *see also* Composite index numbers
- consumer price index, 447-51, 456
 - defined, 427, 455
- industrial production index, 453-55, 457
 - kinds of, 428-30
 - normality of base period, 441, 456
 - price indexes, 429, 430
 - purchasing power index, 430
 - purpose of, 439, 456
 - quantity indexes, 430
 - revisions of, 444-47, 456
 - selection of sample, 439-41, 456
 - simple, 431, 432, 455
 - sources of, 429
 - splicing two series, 446, 447, 456
 - statistical adjustments in, 444
 - substitution of items, 444, 445, 456
 - tests of, 438-44, 456
 - trustworthiness of data, 441, 456
 - value indexes, 430
 - weights, 442
 - bias due to, 443, 444, 456
 - constant or variable, 443
 - physical quantities or values as, 442, 443
 - selection of, 442
 - wholesale price index, 451-53, 456
- Induction, 2
- Industrial production index
 - defined, 453
 - limitation of, 455
 - uses of, 454, 457
- Industrial Revolution, 2
- Industry Surveys*, 648
- Information Please Almanac*, 19
- Interval estimate, 249, 271
- Inventory model, 399-401, 414
 - continuous functions, 402, 403
 - discrete functions, 401, 402
 - goodwill costs, 403, 414
 - scrap allowances, 403, 414
- Irregular fluctuations; *see also* Cyclical fluctuations
 - cause of, 533, 534, 546
 - defined, 546
 - electronic computer for measurement, 538, 539
 - factors in, 534, 546

J

- Joint Economic Committee, 7
- Joint probability, 142-44, 157
- Journal of Commerce, The*, 20
- Judgment sampling, 343, 344, 347

K

Kelvin, Lord, 1
Kurtosis, 114-16

L

Law of growth, 466, 467, 481, 495
Logarithms, 689-91
defined, 1

M

Mail questionnaires, 29, 30
preparation of, 30-32, 37
Management Record, 118
Mantissa, 690
Marginal probability, 142-44, 157
Matrix, defined, 620
Matrix operations, 620-27
multiple regression analysis, 627-30
Mean deviation, 120-22
characteristics of, 132
formulas for, 135, 136
Median
characteristics of, 108
defined, 99
formula, 109
graphic interpolation, 102
grouped data, 100-102
open-end distributions, 101, 102
ungrouped data, 99, 100
Million Random Digits, A, 253
Mode
characteristics of, 108, 109
defined, 102, 103
formula, 109
grouped data, 103, 104
ungrouped data, 103
Modified means
characteristics of, 108
defined, 102
formula, 109
Monte Carlo method
continuous distributions, 415-17
simulation of probability distributions,
411-13, 415
*Monthly Catalog of United States Govern-
ment Publications*, 20
Monthly Labor Review, 19
Monthly Report on the Labor Force, 268
Moroney, M. J., 8
Multiple correlation; *see also* Simple corre-
lation
coefficient of, 600, 601, 612
defined, 551, 552
use of, 589
Multiple regression; *see also* Regression
analysis
basic assumptions in use of, 610, 612
beta coefficients, 603, 612
cautions in use of, 610-12

Multiple regression—*Cont.*

coefficients, estimation of, 593
graphic analysis, 593-97, 611
least squares, 597-99, 611, 612
standard error of estimate, 599, 600,
612
colinearity, 610-12
curvilinear regression, 632-35
defined, 611
electronic computer programs, use of,
603-10, 612
interpretation of results of, 602-10
matrix solution to, 627-30
statistical inference in, 601, 602
standard error of regression coeffi-
cient, 601, 602
time series correlation, 642, 651
use of, 590-93, 611

N

National Bureau of Economic Research, 7
National Industrial Conference Board, 7
New York Times, 430
New York Times Index, The, 20
Nielsen Company, A. C., 30
Noncomparable data, 11
Nonprobability sampling
defined, 340, 341, 346
judgment sampling, 343, 344, 347
quota sampling, 341-43, 346
standard errors of, 344, 348
Normal curve
areas under, 706
table of, 177-79
Normal distribution
approximation to binomial distribution,
179-82
Bayes' theorem for; *see* Bayes' theorem
decision-making, 234-40, 242
defined, 184
normal probability paper, 182-84
proportions, 304, 305, 311
purposes of, 176, 177
table of areas under normal curve, 177-
79
uses of, 176, 177, 184
Normal probability paper, 182-84
Null hypothesis, 291-93, 296
proportions, 311

O

Ogive, 82, 83, 88
Original statistical data
census, 27
collection of, 26 ff.
editing schedules, 33
electronic data processing, 35, 36
follow-up procedure, 32, 33
hand tabulation, 33

Original statistical data—*Cont.*
 mail questionnaires, 29, 30
 preparation of, 30–32, 37
 personal interviews, 28
 preliminary tabulation, 33–36
 punch cards, 33–35
 sampling, 27
 basic reasons for use of, 27, 28
 surveys for collection of, 26, 27

P

Pacific Telephone and Telegraph Company,
 5
 Parameter, 249, 271
 Payoff table, 194, 195
 Percentages, errors in, 14, 15
 Personal interviews, 28
 Point estimate, 249, 271
 Poisson distribution
 approximation to binomial distribution,
 175, 176
 assumptions underlying, 173, 174, 184
 cumulative terms, 725–27
 examples of, 174, 175
 formula of, 173
 individual terms, 722–23
 tables of, 175
 use of, 173, 184
 Population measure, 249
 Posterior distribution, 357, 358, 363–67,
 371, 377–84, 391
Predictions, 524, 545, 648
 President's Council of Economic Advisers, 7
 Price deflation, 471–73, 496
 Price indexes, 429, 430
 Prior distribution, 357–62, 371, 378, 379,
 390
 Probabilities; *see also* Probability distribu-
 tions
 addition of, 146, 147, 158
 basic concepts, 140–46, 157
 brand loyalty, 149–51
 conditional, 142–44, 157
 decision-making based upon, 193, 194
 defined, 140, 157
 estimation of, 141, 142, 157
 examples in use of, 148–53
 joint, 142–44, 157
 marginal, 142–44, 157
 models for; *see* Probability models
 multiplication of, 147, 148, 158
 mutually exclusive events, 146, 147, 158
 project scheduling, 151, 152
 relative frequency of past events, 141,
 157
 rolling dice, 148, 149
 rules for dealing with, 146–48
 sampling, 149
 simple, 142–44, 157

Probabilities—*Cont.*
 simulation, 141
 sources of, 141, 142
 statistical independence, 144–46, 157
 subjective judgment, 142, 157
 decision-making and, 199–202
 theoretical distributions, 141, 142
 Probability distributions; *see also* Binomial
 distribution; Normal distribution;
 Poisson distribution; *and* Probabilities
 Bayes' theorem for revision of; *see* Bayes'
 theorem
 continuous, 154, 155, 158
 defined, 153, 158
 discrete, 154, 155, 158
 expected value of, 156–58
 graphs of, 155
 models for decision-making; *see* Prob-
 ability models
 Monte Carlo method for simulating,
 411–13, 415
 random variable, 153, 158
 simulation, 407–15
 variance of, 156–58
 Probability models, 190, 191
 bidding model, 397–99, 414
 complex systems, simulation of, 413, 414
 inventory model, 399–403, 414; *see also*
 Inventory model
 Monte Carlo method, 411–13, 415–17
 queuing model, 403–7, 414, 415
 simulation, 407–15
 Probability sampling
 cluster sampling, 334–39, 346; *see also*
 Cluster sampling
 defined, 316, 317, 345
 ratio estimation, 327–34, 345, 346; *see*
also Ratio sampling
 replicated sampling, 339, 340, 346
 simple random sampling, 317, 318, 345
 stratified sampling, 319–27, 345; *see also*
 Stratified sampling
 systematic selection, 318, 319, 345
 Probability theory; *see also* Probabilities
 defined, 140
 Problems
 averages, 109–13
 Bayes' theorem
 normal distributions, 392–96
 probability distributions, 372–76
 binomial distribution, 184–88
 charts, effective use of, 62–65
 collection of data, 37–41
 curvilinear regression, 652–57
 cyclical fluctuations, 547–49
 decision-making, 217–23, 242–47
 dispersion, 136–39
 frequency distributions, 88–93
 index numbers, 457–62

Problems—*Cont.*

- irregular fluctuations, 547–49
 - multiple correlation and regression, 612–20
 - normal distribution, 184–88
 - Poisson distribution, 184–88
 - probabilities, 160–65, 217–23
 - probability distributions, 160–65
 - probability models, 417–25
 - proportions, 311–14
 - quality control, 678–81
 - ratios, 88–93
 - regression analysis, 584–88
 - sample survey methods, 348–55
 - seasonal variation, 527–31
 - secular trend, 497–501
 - simple correlation, 584–88
 - small samples, 311–14
 - statistical inference, 272–75
 - statistical quality control, 678–81
 - statistics in business and economics, 16
 - tables, effective use of, 62–65
 - tests of hypotheses, 296–99
 - time series correlation, 652–57
- Process control, 678
- fixed diameter of the disc, 668–72
 - sources of variation, 668
- Project scheduling, 151, 152
- Proportions, 303, 310
- binomial versus normal distribution, 304, 305, 311
 - confidence interval for, 306, 307, 311
 - defined, 304
 - null hypothesis, 311
 - simple random sampling, 318
 - standard error of, 305, 311
 - test of a hypothesis for, 307, 308, 311
 - test of difference between two, 308–11
- Pulse, 30
- Punch cards, 33–35
- Purchasing power index, 430

Q

Quality control

- acceptance sampling, 675, 676, 678
 - types of plans of, 676, 677
- application of methods of, 659
- assignable variation, 660, 677
- chance variation, 660, 677
- chart for averages, 661–63
 - use of, 663–65, 677
- charts for variables, 660–75, 677
- control of attributes, 672–75, 678
- defined, 677
- fraction defective chart, 673–75
 - number of defects per unit, 675
- p* chart, 673–75
- process control, 668–72, 678; *see also* Process control

Quality control—*Cont.*

- quality characteristic, 660
 - R* charts, 663
 - use of, 665, 677
 - ranges of samples, 663
 - use of charts for, 665, 677
 - specifications, 666–68
 - 3-sigma limits for charts of, 665, 666, 677
 - types of variation in quality, 659, 660, 677
 - \bar{X} charts, 661–63
 - use of, 663–65, 677
- Quality Control and Industrial Statistics*, 676
- Quantity indexes, 430, 436
- Quartile deviation, 118–20
- characteristics of, 131, 132
 - formula for, 135
- Queuing model, 403, 414, 415
- characteristics of problems involved, 404
 - one-channel, 404–7
 - simulation of situation, 408–11
 - two-channel, 407
- Quota sampling, 341–43, 346

R

- Random numbers, 730, 731
- table of, 253
 - how to use, 253, 254
- Random sampling, 252–54, 271
- Random variable, 153, 158
- Range
- characteristics of, 131
 - defined, 117
 - formula for, 135
- Ratio charts
- calculations on, 57
 - comparison between two curves, 55, 56
 - constant rate of growth as straight line, 54, 55
 - limitations of, 57, 62
 - nature of, 51–53, 61
 - plotting of, 53, 54
 - uses of, 54–57, 61
- Ratio estimation; *see* Ratio sampling
- Ratio sampling, 327–30, 345, 346
- bias and, 332
 - error associated with, 332–34
 - formulas, 330–32, 347
 - notation of, 330–32
- Ratios
- base of, selection of, 68–70
 - cautions in use of, 70, 71
 - defined, 68, 87
 - denominator of, selection of, 68–70
 - function of, 68
 - numerator of, selection of, 68, 69
- Reciprocals, 695 ff.

Regression, defined, 551
 Regression analysis; *see also* Multiple regression
 coefficient, 554, 567, 568, 582
 sampling error of, 566, 567
 control, 578, 579, 581
 curve fitting, 554-56
 graphic method, 557, 558, 582
 least squares, 558-61, 582
 curvilinear regression, 631, 632
 curvilinear relationships, 579, 580
 examples of, 577-79
 individual forecast, 569-71
 line of average relationship, 556, 568
 prediction, 577, 578, 581
 sampling and, 563
 basic assumptions, 564-66
 confidence intervals, 567-71
 sampling error of regression coefficient, 566, 567
 standard error of estimate, 561-63, 582
 standard error of forecast, 569, 583
 Regression coefficient, 554, 567, 568, 582
 sampling error of, 566, 567
 Regression fallacy, 580, 581
 Relationships between variables; *see* Multiple correlation; Regression analysis; *and* Simple correlation
 Relative dispersion, 132, 133
 formula for, 136
 Replicated sampling, 339, 340, 346
 formula, 347
 Research sources
 accuracy, judging of, 23-25
 cross references, 22, 23
 discrepancies, checking for, 21-23
 importance of, 18, 19, 36
 revisions, 22
 rifle method, 19
 shotgun method, 18
 significant figures in computations, 25, 26
 steps in finding, 19-21
 typographical errors, 22

S

Sample means, confidence intervals for
 difference between, 293-95
 Sample means distribution, 254, 255
 concepts of, 258, 259
 illustration of, 255, 256
 three distributions involved, 256-58
 true mean, estimator of, 260
 Sampling, 27
 basic reasons for widespread use of, 27, 28
 bias
 manner in which sample is taken, 250, 251

Sampling—*Cont.*
 bias—*Cont.*
 measurement inaccuracy, 251
 nonresponse, 251
 cluster; *see* Cluster sampling
 error in, 250
 expected value of information
 Bayes' theorem, 367-70, 372
 factors influencing, 387, 391
 measurement of errors in, 135
 nonprobability, 340-44, 346; *see also*
 Nonprobability sampling
 probabilities used in, 149
 probability, 316-40, 345; *see also*
 Probability sampling
 ratio estimation; *see* Ratio sampling
 regression analysis, 563
 basic assumptions, 564-66
 confidence intervals, 567-71, 583
 sampling error of regression coefficient, 566, 567
 simple random, 252-54, 271
 size, determination of, 268-70
 small; *see* Small samples
 stratified; *see* Stratified sampling
 survey methods, 316 ff., 345
 use of information from, 250
 Sampling error
 correlation coefficient, 576, 577
 regression coefficient, 566, 567
 Sampling interval, 318
 Scatter diagrams, 552-54, 581
 Sciences, statistical analysis in, 1-3, 15
 Seasonal indexes, 523-25
 Seasonal rhythm, 503
 Seasonal variation
 adjustments, 503-5
 calendar variation, 503-5, 525
 changing seasonality, measurement of, 516-18, 526
 daily rhythms, 505
 definition, 507, 525
 electronic computers for measurement of, 518-21, 526
 features of, 503
 graphic method of measurement, 508-12, 526
 revision for greater accuracy, 512
 impact of, 502
 indexes in short-term forecasting, 523-27
 kinds of, 502
 methods of measuring, 507-23, 525
 moving-average method of measurement, 512-16, 526
 purposes of measuring, 506, 525
 selection of method of measurement, 521, 526
 short-term forecasting, 523-27
 weekly rhythms, 505

- Secular trend
 application of, 465
 defined, 465, 495
 graphic "freehand" measurement, 475, 496
 curves, fitting and projecting of, 476–78
 eliminating trend, 478, 479
 group averages, 475
 mathematical methods compared, 479–81
 growth curves, 481–83, 496
 law of growth, 466, 467, 481, 495
 least squares, 483–88, 496
 logarithmic straight line, 490–97
 parabola, 489, 490, 497
 methods of measuring, 473–97
 period of years selected, 469–71, 495
 price deflation, 471–73, 496
 purposes of measuring, 468, 469, 495
 types of, 466, 467
- Selected Business Reference Sources*, 20
- Semantics, errors in, 11, 12
- Sequential decisions, 202–11, 213; *see also* Decision trees
- Sequential sampling, 676, 677
- Simple correlation; *see also* Multiple correlation
 causation, 580
 coefficient of, 571–74
 graphic analysis, 574–76
 sampling error of, 576, 577
 curvilinear, 553, 554, 581
 defined, 551, 581
 linear, 553, 581
 negative, 552, 553, 581
 positive, 552, 553, 581
 scatter diagrams, 552–54, 581
 zero, 552
- Simple index numbers, 431, 432, 455
- Simple probability, 142–44, 157
- Simple random sampling, 252–54, 271, 317, 318, 345
 formula, 347
- Simulation, 141
- Single-sampling plan, 676
- Skewness, 114–16
 coefficient of, 134
 defined, 133
 formula for, 136
- Small samples
 confidence intervals, 302
 normal population, 301, 310
 tests of hypotheses, 302, 303
- Square roots, 694, 695
- Standard deviation, 122–28
 characteristics of, 132
 formulas for, 135, 136
 units of, 133
- Standard error of estimate
 curvilinear regression, 638, 650
 regression analysis, 561–63, 582
- Standard error of nonprobability sampling, 344, 348
- Standard error of the mean, 260–65, 271
Standard Industrial Classification Manual, 22
- Statistical Abstract*, 19, 21, 44
- Statistical analysis
 basic steps in, 15
 control of business operations through, 5
 defined, 1, 15
 economics, use in, 7
 importance of, 1
 scientific method, as, 1–3, 15
 usefulness of, 2
- Statistical and Review Issues of Trade and Business Periodicals*, 20
- Statistical data
 analysis of; *see* Frequency curves; Frequency distributions; *and* Ratios
 assuming causation from correlation, 12
 assumption of stability in a changing economy, 13, 14
 attributes of, 67
 bias, 9
 cautions in use of, 8–16
 classification methods, 42, 43
 collection of; *see* Collection of data
 computation figures, 25, 26
 criteria used in classification of, 67
 decision-making function, 3
 defined, 1
 errors in percentages, 14, 15
 errors in semantics, 11, 12
 faulty deduction, 10, 11
 faulty generalization, 9, 10
 importance of, 1
 misuses of, 8–16
 noncomparable data, 11
 original, 26–36; *see also* Original statistical data
 oversimplification, 12, 13
 presentation of; *see* Charts *and* Tables
 spurious accuracy, 13
- Statistical independence, defined, 144–46, 157
- Statistical inference
 Bayesian approach; *see* Bayes' theorem
 classical approach; *see* Tests of hypotheses
 confidence interval; *see* Confidence intervals
 curvilinear regression, 639
 defined, 2, 249, 271
 multiple regression, 601, 602
 standard error of regression coefficient, 601, 602

Statistical inference—*Cont.*
 proportions, 303–11; *see also* Proportions
 small samples, 300–303, 310
 tests of hypotheses; *see* Tests of hypotheses
 Statistical methods; *see* Statistical analysis
Statistical Methods in Quality Control, 676
Statistical Quality Control, 676
 Statistical quality control; *see* Quality control
Statistical Services of the United States Government, 20
Statistical Yearbook, 19
Statistics, 427
 Statistics; *see also* Statistical analysis
 business problem solution by, 4–6
 decision-making, function in, 3, 4
 defined, 249, 271
 economic research, uses in, 6–8
 elements of, 1
 Stratification, 319, 345
 Stratified sampling
 allocation of sample to strata
 least-cost, 326, 327, 345
 optimum, 324–26, 345
 proportional, 323, 324, 345
 defined, 319, 345
 estimate of the mean and standard error, 321–23
 example of, 320, 321
 formulas, 347
 heterogeneous populations, 320
 homogeneous populations, 320
 nonresponse, 327
 Subjective probability, 142, 157
 decision-making and, 199–202
Survey of Current Business, 19, 427
 Systematic sampling, 318, 319, 345

T

Tables
 advantages over charts, 42, 60
 arrangement of, 44–46
 classifications used in, 42, 43, 60
 construction of, 44–46, 61
 defined, 42
 random numbers, 253, 254
 reference, 44
 summary, 44
 types of, 43, 44
 Tests of hypotheses
 acceptance of hypothesis, 277, 278
 arithmetic means, differences between, 289–94
 choice between acceptance and rejection of hypothesis, 280, 281
 confidence intervals for difference between sample means, 293–95

Tests of hypotheses—*Cont.*
 critical probability, 281, 294
 errors in, 281
 balancing one against other, 285, 286
 sample size affecting, 286, 295
 type I, 281–83, 294
 type II, 283–85, 294
 example of approach, 276–81, 294
 level of significance, 281
 null hypothesis, 291–93, 296
 one-tailed, 288, 289, 295
 operating characteristic curves, 283–85, 295
 proportions, 307, 308, 311
 rejection of hypothesis, 278, 279
 sample size as affecting probability of errors, 286, 294
 small samples, 302, 303
 two-tailed, 287, 288, 295
 use of, 276, 294
 Time series analysis; *see* Business fluctuations, types of
 Time series correlation, 640
 actual annual data, 641–45, 651
 first differences, 641, 645, 646, 651
 methods of, 640–46, 651
 multiple regression, 642, 651
 percents of trend, 641, 642, 651
 Trend fitting, sums of squares and fourth powers used in, 729
Trendex, 30
TV Guide, 30
 Twentieth Century Fund, 7

U

Unit normal loss function, 707
United States Department of Commerce Publications, 20
 Utility curve for decision-making, 214–17

V

Value indexes, 430
 Values of t , 728
 Variables, 67; *see also* Frequency distributions
 Variances
 continuous distribution, 159
 defined, 123
 probability distributions, 156–58

W

Waiting lines, 403–7, 414, 415
Wall Street Journal Index, The, 20
 Weighted mean, 96
 Wholesale price index
 basis of, 451
 defined, 451
 limitations of, 453
 uses of, 452, 453, 456
World Almanac, The, 19

This book has been set in 12 and 10 point Garamond #3, leaded 1 point. Chapter numbers and titles are in 18 point Lydian Bold; part titles are in 18 point Lydian caps. The size of the type page is 27 x 46½ picas.
